Multimodal transformers for emotion recognition

Juan Fernando Vazquez Rodriguez

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Informatique
Unité de recherche : Laboratoire d'Informatique de Grenoble

Transformateurs multimodaux pour la reconnaissance des émotions

Multimodal transformers for emotion recognition

Présentée par :

Juan Fernando VAZQUEZ RODRIGUEZ

Direction de thèse :

James CROWLEY PROFESSEUR DES UNIVERSITES EMERITE, GRENOBLE INP	Directeur de thèse
Grégoire LEFEBVRE DOCTEUR EN SCIENCES, ORANGE INNOVATION	Co-encadrant de thèse
JULIEN CUMIN DOCTEUR EN SCIENCES, ORANGE INNOVATION	Co-encadrant de thèse

Rapporteurs :

MOHAMED CHETOUANI
PROFESSEUR DES UNIVERSITES, SORBONNE UNIVERSITE
BJORN W. SCHULLER
FULL PROFESSOR, IMPERIAL COLLEGE LONDON

Thèse soutenue publiquement le **27 novembre 2023**, devant le jury composé de :

JAMES CROWLEY PROFESSEUR DES UNIVERSITES EMERITE, GRENOBLE INP	Directeur de thèse
MOHAMED CHETOUANI PROFESSEUR DES UNIVERSITES, SORBONNE UNIVERSITE	Rapporteur
BJORN W. SCHULLER FULL PROFESSOR, IMPERIAL COLLEGE LONDON	Rapporteur
PATRICK REIGNIER PROFESSEUR DES UNIVERSITES, GRENOBLE INP	Président
JUSTINE CASSELL FULL PROFESSOR, CARNEGIE MELLON UNIVERSITY	Examinatrice

Invités :

GREGOIRE LEFEBVRE
DOCTEUR EN SCIENCES, ORANGE INNOVATION
JULIEN CUMIN
DOCTEUR EN SCIENCES, ORANGE INNOVATION

ABSTRACT

Mental health and emotional well-being have significant influence on physical health, and are especially important for healthy aging. Continued progress on sensors and microelectronics has provided a number of new technologies that can be deployed in homes and used to monitor health and well-being. These can be combined with recent advances in machine learning to provide services that enhance the physical and emotional well-being of individuals to promote healthy aging. In this context, an automatic emotion recognition system can provide a tool to help assure the emotional well-being of frail people. Therefore, it is desirable to develop a technology that can draw information about human emotions from multiple sensor modalities and can be trained without the need for large labeled training datasets.

This thesis addresses the problem of emotion recognition using the different types of signals that a smart environment may provide, such as visual, audio, and physiological signals. To do this, we develop different models based on the Transformer architecture, which has useful characteristics such as their capacity to model long-range dependencies, as well as their capability to discern the relevant parts of the input. We first propose a model to recognize emotions from individual physiological signals. We propose a self-supervised pre-training technique that uses unlabeled physiological signals, showing that that pre-training technique helps the model to perform better. This approach is then extended to take advantage of the complementarity of information that may exist in different physiological signals. For this,

we develop a model that combines different physiological signals and also uses self-supervised pre-training to improve its performance. We propose a method for pre-training that does not require a dataset with the complete set of target signals, but can rather, be trained on individual datasets from each target signal.

To further take advantage of the different modalities that a smart environment may provide, we also propose a model that uses as inputs multimodal signals such as video, audio, and physiological signals. Since these signals are of a different nature, they cover different ways in which emotions are expressed, thus they should provide complementary information concerning emotions, and therefore it is appealing to use them together. However, in real-world scenarios, there might be cases where a modality is missing. Our model is flexible enough to continue working when a modality is missing, albeit with a reduction in its performance. To address this problem, we propose a training strategy that reduces the drop in performance when a modality is missing.

The methods developed in this thesis are evaluated using several datasets, obtaining results that demonstrate the effectiveness of our approach to pre-train Transformers to recognize emotions from physiological signals. The results also show the efficacy of our Transformer-based solution to aggregate multimodal information, and to accommodate missing modalities. These results demonstrate the feasibility of the proposed approaches to recognizing emotions from multiple environmental sensors. This opens new avenues for deeper exploration of using Transformer-based approaches to process information from environmental sensors and allows the development of emotion recognition technologies robust to missing modalities. The results of this work can contribute to better care for the mental health of frail people.

Keywords Transformers, Emotion Recognition, Self-Supervised Learning, Pre-Training, Machine Learning.

RÉSUMÉ

La santé mentale et le bien-être émotionnel ont une influence significative sur la santé physique, et sont particulièrement importants pour un viellissement en bonne santé. Les avancées continues dans le domaine des capteurs et de la microélectronique en général ont permis l'avènement de nouvelles technologies pouvant être déployées dans les maisons pour surveiller la santé et le bien-être des occupants. Ces technologies de captation peuvent être combinées aux avancées récentes sur l'apprentissage automatique pour proposer des services utiles pour vieillir en bonne santé. Dans ce cadre, un système de reconnaissance automatique d'émotions peut être un outil s'assurant du bien-être de personnes fragiles. Dès lors, il est intéressant de développer un système pouvant déduire des informations sur les émotions humaines à partir de modalités de captation multiples, et pouvant être entrainé sans requérir de larges jeux de données labellisées d'apprentissage.

Cette thèse aborde le problème de la reconnaissance d'émotions à partir de différents types de signaux qu'un environnement intelligent peut capter, tels que des signaux visuels, audios, et physiologiques. Pour ce faire, nous développons différents modèles basés sur l'architecture *Transformer*, possédant des caractéristiques utiles à nos besoins comme la capacité à modéliser des dépendances longues et à sélectionner les parties importantes des signaux entrants. Nous proposons en premier lieu un modèle pour reconnaitre les émotions à partir de signaux physiologiques individuels. Nous proposons une technique de pré-apprentissage auto-supervisé utilisant des données

physiologiques non-labellisées, qui améliore les performances du modèle. Cette approche est ensuite étendue pour exploiter la complémentarité de différents types de signaux physiologiques. Nous développons un modèle qui combine ces différents signaux physiologiques, et qui exploite également le pré-apprentissage auto-supervisé. Nous proposons une méthode de pré-apprentissage qui ne nécessite pas un jeu de données unique contenant tous les types de signaux utilisés, pouvant au contraire être pré-entrainé avec des jeux de données différents pour chaque type de signal.

Pour tirer parti des différentes modalités qu'un environnement connecté peut offrir, nous proposons un modèle multimodal exploitant des signaux vidéos, audios, et physiologiques. Ces signaux étant de natures différentes, ils capturent des modes distincts d'expression des émotions, qui peuvent être complémentaires et qu'il est donc intéressant d'exploiter simultanément. Cependant, dans des situations d'usage réelles, il se peut que certaines de ces modalités soient manquantes. Notre modèle est suffisamment flexible pour continuer à fonctionner lorsqu'une modalité est manquante, mais sera moins performant. Nous proposons alors une stratégie d'apprentissage permettant de réduire ces baisses de performances lorsqu'une modalité est manquante.

Les méthodes développées dans cette thèse sont évaluées sur plusieurs jeux de données. Les résultats obtenus montrent que nos approches de *Transformer* pré-entrainé sont performantes pour reconnaitre les émotions à partir de signaux physiologiques. Nos résultats mettent également en lumière les capacités de notre solution à aggréger différents signaux multimodaux, et à s'adapter à l'absence de l'un d'entre eux. Ces résultats montrent que les appproches proposées sont adaptées pour reconnaitre les émotions à partir de multiples capteurs de l'environnment. Nos travaux ouvrent de nouvelles pistes de recherche sur l'utilisation des *Transformers* pour traiter les informations de capteurs d'environnements intelligents et sur la reconnaissance d'émotions robuste dans les cas où des modalités sont manquantes. Les résultats de ces travaux peuvent contribuer à améliorer l'attention apportée à la santé mentale des personnes fragiles.

Mots-clés Transformers, reconnaissance d'émotions, apprentissage auto-supervisé, pré-apprentissage, apprentissage automatique.

ACKNOWLEDGEMENTS

I would like to express my immense gratitude to my advisors Prof. James, Grégoire, and Julien, who have been of immense help and without whom completing this thesis would have been impossible. I feel lucky to have as advisors researchers who not only helped conduct my research with scientific rigor, but also helped me to navigate the day-to-day life of a Ph.D. candidate. I thank them for all their help and dedication. I would also like to thank Mohamed Chetouani and Bjoern Schuller for reviewing this manuscript, as well as the rest of the members of the jury for evaluating my thesis.

I deeply thank my wife Anita, who with extraordinary patience and love cared for me through these years. Her support has been fundamental during this thesis. Without her, I wouldn't even be in France, let alone finishing a thesis.

I would also like to thank my coworkers at Orange. Erwan, the manager of my team, who has always been attentive to what I need to do my work. Hervé, who was always helpful. Pauline, who introduced me to the board game world inside Orange and with whom I could always count to answer my questions about life in France. I also thank Xi and Miriam, who sooner than they think will be in my position. Also, I want to express my gratitude to the rest of my colleagues (and ex-colleagues) who have made Orange a nice place to work. Thanks also to the members of the M-PSI team from LIG at the Université Grenoble Alpes, especially Sannara, Louis, and Yangtao,

and to the team leader, Dominique, who welcomed me there.

I also thank to my family, my parents Fanny and Paciente, who raised me as a curious person, without fears of trying new adventures such as leaving the country to pursue new opportunities. I feel privileged to have them as parents. My sister Natalia, who has always been there for me, even coming all the way from the Netherlands to spend time together, supporting Grenoble's heat waves. And my cousin Andres, with whom I spend fantastic hours of virtual entertainment, and spend even more amazing hours of real friendship. To my friends from Ecuador, Pablo, Andres (Nine), Pedro, Cristina, and Cristian, thanks for your support.

To everyone, simply: Gracias!

Juan

CONTENTS

NOMENCLATURE

Abbreviations

1D-CNN	1D Convolutional Neural Network. 47, 49, 60, 61, 67, 75, 91, 114
2D-CNN	2D Convolutional Neural Network. 48, 75, 104
AMMTD	Autoregressive Multimodal Transformer Decoder. xi, xxi, xxiv, 120, 123–126, 134–137, 139, 152, 154, 155, 160–162, 166
AU	Action Unit. 18, 130, 158, 183
BOHB	Bayesian Optimization with Hyperband. 68, 69, 99
BPM	Beats per Minute. 130, 158, 182, 183
CCC	Concordance Correlation Coefficient. xxi, xxii, 128, 129, 131, 132, 138, 158–165, 170, 171
CNN	Convolutional Neural Network. 30, 46, 47, 49, 51, 56, 57, 64, 75, 121, 183
DCCA	Deep Canonical Correlation Analysis. 104
DL	Deep Learning. x, xx, 5, 29, 30, 41–43, 45–49, 51, 52, 57, 64, 82, 84–90, 92, 93, 104, 105, 144, 168, 169

ECG	Electrocardiogram. xix, xx, xxiii, xxiv, 17, 21, 22, 25, 26, 36, 40, 41, 43, 44, 47, 49, 52, 55, 57, 65, 66, 68–77, 82, 83, 88, 89, 91, 95, 97–101, 104, 105, 130, 158, 175, 179–183, 185–187
EDA	Electrodermal Activity. xix, 17, 23–25, 43, 44, 47, 88, 89, 91, 104, 181, 182, 184
EDR	Electrodermic Response. 24
EEG	Electroencephalogram. xix, xx, xxiv, 17, 22–26, 29, 36, 40–45, 47–50, 52, 55, 57, 65, 66, 68, 69, 72, 73, 75, 76, 82, 83, 88, 89, 91, 92, 95, 97–101, 104, 105, 179–181, 185–187
eGeMAPS	extended Geneva Minimalistic Acoustic Parameter Set. 20, 130, 158, 182, 183
EMG	Electromyogram. 17, 88
ERN	Emotion Regression Network. 120, 124, 127, 131
FACS	Facial Action Coding System. 18, 19, 183
FCN	Fully-Connected Network. xxii, xxiv, 33, 34, 46–49, 56, 57, 62, 64, 68, 69, 71, 72, 87, 88, 96, 99, 102, 104, 114, 115, 120, 127, 136, 137, 139, 149, 150, 160–162, 170
FFN	Feed-Forward Network. 62, 68, 124–126, 131
GAN	Generative Adversarial Networks. 148, 151
GELU	Gaussian Error Linear Units. 131
GeMAPS	Geneva Minimalistic Acoustic Parameter Set. 19
GNB	Gaussian Naive Bayes. 86, 89, 104
GRU	Gated Recurrent Unit. 30, 104
LSTM	Long Short-Term Memory. xxi, xxii, xxiv, 30, 47, 75, 104, 112–116, 133–137, 139, 160–162, 170
MHA	Multi-Head Attention. 32–34, 61, 62, 124–126
MHCA	Multi-Head Cross-Attention. 123–127, 137
MHSA	Multi-Head Self-Attention. 123–125
ML	Machine Learning. 5, 28, 29, 43, 45, 47, 48, 83, 86, 87, 89, 178
MMTE	Multimodal Transformer Encoder. xi, xxi, xxii, xxiv, 120–122, 126, 127, 134, 136, 137, 152–155, 159, 161, 162, 166
MSE	Mean Square Error. xxi, 132

NLP Natural Language Processing. 30, 46, 49, 51, 59

PCA Principal Component Analysis. 89
PPG Photoplethysmography. 88, 89

RAAW Rater Aligned Annotation Weighting. 184
RBM Restricted Boltzmann Machine. 91, 92
ReLU Rectified Linear Unit. 67–69, 99, 131
RESP Respiration. 17, 88, 130, 158, 182, 183
RMSE Root Mean Square Error. xxii, 128, 129, 134, 135, 137, 138, 159–165, 170, 171
RNN Recurrent Neural Network. xxi, 30, 46, 47, 64, 112–114, 119

SVM Support Vector Machine. 29, 43–45, 75, 86, 89, 92, 104

TCN Temporal Convolutional Network. 121, 122, 131
TDL Transformer Decoder Layer. 34, 123–126
TEMP Temperature. 17
TSST Trier Social Stress Test. 130, 182

ULM-TSST Ulm-Trier Social Stress Test. xiii, 8, 37, 129, 130, 133, 135, 158, 166, 182–184

VAE Variational Autoencoder. 91, 104

WHO World Health Organization. 1, 3

List of figures

List of tables

CHAPTER 1

INTRODUCTION

Mental health plays an important role in having a healthy life. According to the World Health Organization (WHO), mental health is "a state of mental well-being that enables people to cope with the stresses of life, realize their abilities, learn well and work well, and contribute to their community" [203]. The WHO also states that mental health is an integral component of health, and it is a basic human right.

An important part of mental health is emotional wellness. Braun and Kloss [27] say the following about emotional wellness:

> "Emotional wellness is when a person's belief of what they are feeling becomes realized into physical manifestations of that belief. For example, feeling encouraged and supported produces positive endorphins... Alternatively, when a person is feeling stressed or overwhelmed, they ostracize or isolate themselves... Often, physical manifestations are associated with poor emotional wellness. Blood pressure problems, heart conditions, and general exhaustion are some common symptoms."

Mental health and emotional well-being have significant influence on physical health, and are especially important for healthy aging. Although mental and emotional health problems affect people from all age ranges, as the world population ages it is becoming increasingly important to pay attention to the mental health of elderly and frail people, especially if these people are living alone. Continued progress in sensors and microelectronics has provided a number of new technologies that can be deployed in homes and used to monitor health and well-being. These can be combined with recent advances in machine learning to provide services that enhance the physical and emotional well-being of individuals.

In this thesis, we aim to develop methods that can be helpful for the general task of preserving the mental and emotional well-being of frail people. To understand better how we do this, it is important to define how we interpret the terms *affect*, *emotion*, and *mood*, especially because there is no consensus in their definition [7].

— ***Affect:*** is a term that covers various feelings that individuals can undergo, encompassing both emotions and moods.

— ***Emotion:*** is a strong short-term feeling usually directed towards a stimulus. Emotions frequently manifest through corporal and facial expressions and bodily responses.

— ***Mood:*** is a mental state milder in intensity than an emotion, which does not necessarily require a specific trigger. Moreover, moods persist for an extended period compared to emotions, lasting from hours to days.

If unattended, negative emotions can turn into negative moods. For this reason, it is important to detect negative emotions before they become more permanent negative feelings. With this, the goal of this thesis is to develop methods for automatically detecting the emotional state of a person. Specifically, we are interested in recognizing if the person is experiencing a positive or negative emotion, and how intense the emotion is. We do not aim to build a system that provides a comprehensive diagnosis and treatment of mental and emotional health issues of a person. Instead, the methods presented in this thesis are intended as a part of a global health system, where other participants like caregivers, family, and maybe additional computerized systems, monitor, diagnose, treat, and help the person.

1.1 Frail People, Emotional Wellness, Smart Environments

We believe that a system capable of recognizing the emotional state of a person can be useful for accommodating the increasing frailty of aging people living alone. As defined by Sicsic et al. [174], frailty is a "geriatric syndrome resulting from age-related cumulative declines across multiple physiologic systems, with reduced capacity of the organism to withstand stress, thus increasing vulnerability to adverse health outcomes including falls, hospitalization, institutionalization, and mortality". Thus, it can be seen that frail people are particularly exposed to the consequences of poor emotional health.

Moreover, the world population is aging. In fact, according to the WHO [204], between 2015 and 2050 the population of older adults is projected to increase from approximately 12% to 22%, which corresponds to an increase from 900 million to 2 billion people older than 60. In addition to this, an important percentage of older people are living alone. For instance, in 2018 in the 27 countries of the European Union, there were 40.2% of women and 21.8% of men aged 65 or more that were living alone [63].

Under this scenario, it could be helpful for a frail person living alone to live in a smart environment, which Kaswan et al. [101] define as an ecosystem equipped with communicating smart objects that could gather information through different sensors and provide services according to the user's needs. Some examples of smart objects in a smart environment include improved traditional appliances, such as refrigerators that monitor and display contents, and order new foodstuff when needed; smart speakers that can provide interactive access to information and media over the internet; and smart wearables, including smartwatches, that can monitor health while providing hand-free mobile access to communications and services. In a smart environment, such objects can be made to work together to provide services to the user [45].

Smart objects can provide a rich ensemble of information about individuals and their environments. For example, there might be sensors to measure temperature, air quality and other ambient variables. The environment can also be equipped with wearables fitted with accelerometers to detect the movements of the user [43], and these wearables might be fitted with sensors that can gather physiological signals. Audio and video might be captured through the interaction with a smart assistant. Such data can

be combined and interpreted to provide a rich source of information for services to enhance quality of life. Machine learning provides the enabling technology for such services.

1.1.1 The Usefulness of Emotion Recognition Systems

We now exemplify how it can be helpful for a frail person to live in a smart environment that has the capacity to recognize emotions when he or she has affective struggles. In this case, the person could be using a wearable that monitors his or her physiological signals, for example, cardiac signals. Through these signals, it can be recognized that the person is not feeling well mentally, perhaps detecting a repetitive negative emotion of high intensity. The smart environment may decide to act [150], initializing an interaction through a camera-equipped smart assistant. Through the audio and video captured by the assistant, the emotion that the person is feeling at that moment can be detected, and this information can be used by the system to further assess his or her affective state. With this assessment, the system can evaluate the next actions. For example, if the affective struggle is not very serious, the system may suggest to the person that he or she should do an activity like taking a walk or doing relaxation exercises. Or maybe the system realizes that help is needed, and communicates the situation to a family member or a caregiver, so they can take any necessary action.

From the previous example, it can be observed that an important component of the described system is being able to recognize the emotional state of the person. In other words, it is necessary to detect how intense it is the emotion that this person is feeling, and how positive or negative this emotion is. Also, note that the smart environment may provide different types of signals, and being capable of processing these different signals to recognize emotions can be advantageous as the information that each type provides might be complementary. Moreover, these sources may not all be available, for example, the user may decide not to activate the camera of the assistant. Therefore, it is desirable that a system that recognizes emotions from different types of signals could keep working when one type of signal is missing.

1.1.2 Practical Considerations for Emotion Recognition

For an emotion recognition system to work inside a system that monitors the global health of a person, it is necessary that this emotion recognition

system works accurately and reliably, which is a challenging problem. Some difficulties are that emotions are subjective feelings, emotions may vary across cultures [119], collecting data related to emotions is difficult [62], bodily response to emotions may vary across individuals [81], and emotions are expressed through different modalities [44], so fusing their information could be necessary. Through this work, we aim to address some of these difficulties and improve the accuracy and reliability of emotion recognition systems, so they can be used as part of a global health system that enhances the care given to frail people.

To summarize this section, we believe that being capable of taking advantage of the signals that a smart environment may provide to perform emotion recognition accurately and reliably is an important problem, that could be especially helpful for frail people living alone. This is why in this thesis we are interested in developing emotion recognition algorithms using different types of input signals, and we also are interested in addressing the problem that sometimes some signals may not be available. Even with the difficulties of gathering data related to emotions, training data is becoming more available, and thus our algorithms are based on Machine Learning (ML), or more specifically, on Deep Learning (DL).

1.2 Contributions of this Thesis

In this thesis, we propose four contributions:

Self-Supervised Learning for Emotion Recognition with Physiological Signals: We propose a method to recognize emotions from single physiological signals, developing a self-supervised technique to pre-train the model to overcome the difficulty that labeled data with labels of emotion is not abundant. The pre-training technique uses unlabeled physiological data, thus it takes advantage of datasets that contain physiological data that have not necessarily been collected for emotion recognition tasks. We argue that using this pre-training technique should improve the accuracy of the model. We evaluate our approach on state-of-the-art datasets. This contribution was published and presented at a peer-reviewed conference as the paper:

Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin and James L. Crowley, "Transformer-Based Self-Supervised Learning for Emotion Recognition". In *26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 2605-2612.

Emotion Recognition from Multiple Physiological Signals: We ex-

tend the contribution described in the previous paragraph and we propose an approach to recognize emotions from multiple physiological signals. This approach uses a self-supervised pre-training technique in which each dataset in the pre-training dataset collection does not need to contain all the concerned physiological signals, but instead different datasets cover one physiological signal each. This is advantageous since finding or collecting datasets that contain all relevant physiological signals is more difficult than obtaining multiple single-signal datasets, even if these datasets do not need to be labeled. We evaluate the accuracy of our approach on state-of-the-art datasets. This contribution was published and presented at a peer-reviewed conference as the paper:

Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin and James L. Crowley, "Emotion Recognition with Pre-Trained Transformers Using Multimodal Signals". In *10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2022, pp. 1-8.

Multimodal Time-Continuous Emotion Recognition: We propose a method to perform time-continuous emotion recognition using multimodal inputs. Our method uses attention mechanisms to aggregate information from the different modalities, and auto-regression to take past predictions into account. We argue that these techniques should be helpful in improving the accuracy of the model. We test our approach on a state-of-the-art dataset. This contribution was part of a paper published and presented at a peer-reviewed conference as:

Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin and James L. Crowley, "Accommodating Missing Modalities in Time-Continuous Multimodal Emotion Recognition". In *11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2023.

Accommodating Missing Modalities in Multimodal Emotion Recognition: We propose an approach capable of performing time-continuous multimodal emotion recognition that is robust to missing modalities. For this, we show that an attention-based model is well adapted for this case, and we also show that the robustness of the model to missing modalities can be improved through a novel training technique. We run several experiments to evaluate our approach on a state-of-the-art dataset. This contribution was published and presented at a peer-reviewed conference, as part of the paper mentioned above.

1.3 Thesis Overview

This manuscript reviews the overall scientific context for our contributions, followed by a presentation of the technological approach, implementation, and experimental evaluation of each contribution, moving from interpretation of sensor information from an individual sensor, combining information from multiple sensors, using this to provide a time continuous multimodal emotion recognition, and describing how our approach can overcome a partial absence of information from individual sensors.

Chapter 2 presents various concepts necessary to address the overall problem of emotion recognition. We provide a review of emotions as studied in the field of Psychology, talking about the definition of emotion, how it can be represented, and the sensing modalities that can be used to capture emotions. We also define what emotion recognition is, and the elements of an emotion recognition system. In addition, we discuss general computing approaches to build an automatic emotion recognition system and review the Transformer architecture[196] in particular, which constitutes the backbone of the approaches developed in this thesis. Finally, we discuss different datasets available for the task of emotion recognition, identifying the datasets used in this thesis to evaluate our contributions.

Chapter 3 presents our contribution to the problem of emotion recognition from a single physiological signal. We start this chapter by giving a definition of the problem being addressed, identifying the challenges associated with solving that problem, and the motivations for why we set up the problem the way we did. After this, we perform a survey of related state-of-the-art contributions. We continue the chapter by introducing our contribution regarding emotion recognition from physiological signals, detailing how we address the different challenges that emerge when addressing this problem. In particular, we give details about our Transformer-based architecture, and we also show how we pre-train that architecture in a self-supervised manner, taking advantage of unlabeled data that is typically more available than labeled data. We end the chapter by showing the evaluation of our approach, which was done by running multiple experiments using AMIGOS [135] and DREAMER [102] datasets, and then we draw conclusions from these experiments.

Chapter 4 presents our contribution related to using multiple physiological signals to perform emotion recognition. At the beginning of this chapter, we define the problem addressed, identifying as one of the main challenges that the size of available labeled data that contains multiple

physiological signals is limited. We then review the state-of-the-art literature that addresses the task of recognizing emotions from multiple physiological signals. Next, we present our contribution, explaining how we pre-train our proposed Transformer-based model capable of processing multimodal physiological signals, to address the challenge identified at the beginning of the chapter. More specifically, we describe a pre-training strategy that allows the use of unlabeled datasets that do not need to contain all the concerned physiological signals at the same time, but instead uses different datasets each containing one of the concerned signals. Finally, we evaluate our approach by running multiple experiments using AMIGOS [135] and DREAMER [102] datasets.

Chapter 5 presents our contribution to the problem of time-continuous multimodal emotion recognition. After providing a definition of the problem addressed in the chapter and explaining the associated challenges, we provide a literature review of the related state-of-the-art contributions. We then present our approach to perform time-continuous multimodal emotion recognition, detailing how we use the attention layers of a Transformer to aggregate the information from the different modalities, and use auto-regression to take into account past predictions. We end the chapter showing different experimental results obtained to test our approach, using the Ulm-Trier Social Stress Test (ULM-TSST) dataset [180].

Chapter 6 presents our contribution to the problem of accommodating missing modalities when performing multimodal emotion recognition. We start this chapter by giving the motivations of why it is pertinent to handle missing modalities in this case. After this, we present relevant state-of-the-art contributions that also address the problem of missing modalities. Next, we detail our approach to accommodating missing modalities in multimodal emotion recognition. We show that our Transformer-based model can handle missing modalities without any architectural change, albeit with some drop in performance. We also describe our training technique designed to make the model more robust to missing modalities, alleviating the problem of performance drop when a modality is missing. Finally, we show experimental results when evaluating our approach with the ULM-TSST dataset [180].

Chapter 7 concludes this thesis, providing a summary of our contributions, and discussing the limitations of our approaches. We also examine some perspectives that emerge from our work, and discuss the potential for ethical misuse and the need for clear ethical guidelines for applications that deal with emotions.

CHAPTER 2

BACKGROUND ON EMOTION RECOGNITION

In this chapter, various concepts and definitions necessary to better understand this thesis are explained. There are four different subjects that are covered, starting in Section 2.1 with some background information about emotions, discussing their definition, and how they can be detected and measured. Then, Section 2.2 provides a definition of emotion recognition and describes how an emotion recognition system might be designed. After this, Section 2.3, describes the Transformer [196], an architecture that constitutes the backbone of the different approaches that are presented in this thesis. Finally, Section 2.4 provides an overview of datasets employed for emotion recognition, identifying the datasets used to evaluate our approaches.

2.1 Emotions

The measurement, and even the definition, of human emotions remains a controversial subject. A reflection of this is the statement of Fehr and

Russell [67] that says: "*Everyone knows what an emotion is, until asked to give a definition. Then it seems, no one knows*". In an attempt to find a definition, Kleinginna and Kleinginna [108] analyzed nearly one-hundred definitions of emotions, and found that they can be classified into eleven categories, based on what they emphasize more. For example, they found that some definitions placed emphasis on feelings of excitement/depression or pleasure/displeasure, others placed emphasis on appraisal processes, some others emphasized the physiological mechanism of emotions, others emphasized emotional and expressive behavior, and other definitions placed emphasis on the functional aspects of emotions. From this analysis, Kleinginna and Kleinginna [108] came up with the following definition:

> "*Emotion is a complex set of interactions among subjective and objective factors, mediated by neural hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal-directed, and adaptive.*"

Although it is difficult to come up with a universal definition of emotion, it can be seen from the work of Kleinginna and Kleinginna [108] that in general, the definitions of emotions involve concepts such as feelings, arousal, pleasure/displeasure, cognitive process, appraisal, physiological responses, and behavior.

2.1.1 Theories of Emotion

There is no general consensus around the internal mechanisms of how an emotion is generated. James [97] and Lange [116] proposed an early theory of emotion, known as the James-Lange theory, that states that emotions are the sensation of bodily changes. James stated that "My theory ... is that the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur IS the emotion".

In contrast to the James-Lange theory, Cannon [33] and Bard [19] presented a theory that states that emotions are generated by a stimulus that is processed by the central nervous system, and physiological reactions are not considered as a cause for emotion elicitation.

A third point of view regarding emotions is the appraisal theory of emotions, which states that an emotion is produced after a cognitive evaluation

of a stimulus (see Scherer and Moore [166]). In this case, the resulting emotion is an outcome of the personal judgment, or appraisal, of an event. This theory explains why people experiencing the same stimulus, watching a moving video for example, will not necessarily experience the same emotion.

Another important aspect of the theory of emotions is the discussion about their universality. According to Kuang et al. [113], "emotion universality theories assume that emotions are innate and universal, independently of human's acknowledgment of them, and that all humans have the capacity to experience and perceive the same core set of emotion categories". Some works in this direction are the works Ekman [55], and Ekman and Friesen [58], who defend that facial expressions of emotion are universal. On the other hand, other authors like Barret [20] and Lindquist et al. [124] affirm that emotions are shaped by societal influences and past experiences, challenging the notion of the universality of emotions.

To summarize, diverse theories have been proposed to explain the mechanism behind emotions, and there is no agreement in the scientific community about how these mechanisms work. However, it can be noted that the presented theories agree that emotions are triggered by a stimulus, and the body displays a reaction associated with the perceived emotion. With these ideas, we continue reviewing emotions by describing how they can be represented.

2.1.2 Emotion Representation

Emotions may be represented by qualitative descriptors that describe how a person feels. Intuitively, a representation of emotion is a word like "*happy*", or saying something like "*In a scale from 1 to 10 describing how positive is the emotion I am feeling, I feel 8*". In fact, these intuitions are the basis of the two main ways of representing emotions: discrete representations and continuous representations. These types of representations are well grounded in models of emotions that come from psychological theories, which are described below.

2.1.2.1 Discrete Representation: Basic Emotion Theory

According to the basic emotion theory, there is a limited number of core or basic emotions, under the assumption that there is a strong agreement in the manner that people express and perceive emotions. They are called discrete because they are distinguishable from one another. One of the most notable proponents of this theory is Paul Ekman [57].

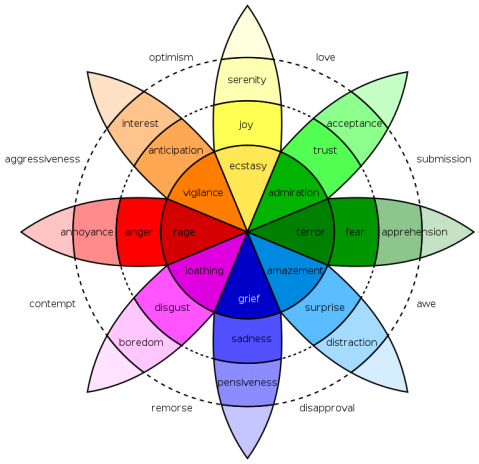
Common examples of discrete emotions include: happiness, fear, disgust, anger, sadness, and surprise, which were used by Ekman et al. in [54]. Nonetheless, researchers do not agree on what constitutes a basic emotion. For instance, Izard [96] lists as basic emotions the following: guilt, shame, contempt, joy, interest-excitement, surprise, distress-anguish, anger, disgust, and fear. In [57], Ekman argued that in order for an emotion to be considered basic, it has to have the following characteristics

1. Distinctive universal signals
2. Distinctive physiology
3. Automatic appraisal
4. Distinctive universals in antecedent events
5. Distinctive appearance developmentally
6. Presence in other primates
7. Quick onset
8. Brief duration
9. Unbidden occurrence
10. Distinctive thoughts, memories images
11. Distinctive subjective experience

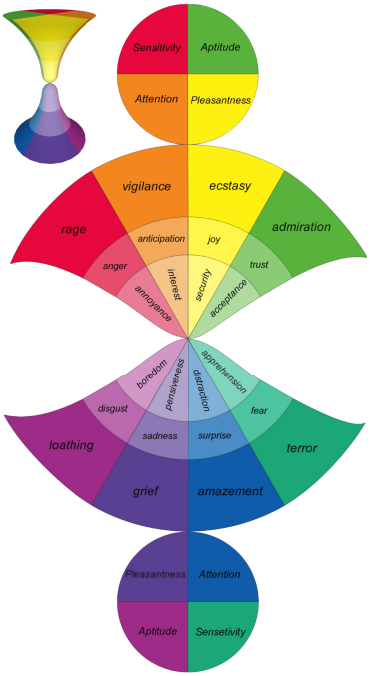
Using those characteristics, Ekman [57] proposes the following list of basic emotions: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure, and shame.

Plutchik [148] proposed the Wheel of Emotions as a way to represent emotions. Plutchik's model, uses 8 basic emotions (joy, trust, fear, surprise, sadness, disgust, anger, and anticipation), mapping them to 3 levels of intensity, as shown in Figure 2.1a. In this model, opposite emotions are placed on opposite sides of the graph; for example, joy is the opposite of sadness. Moreover, the combination of two basic emotions produces a complex emotion, like the combination of joy and anticipation produces optimism and the combination of fear and sadness produces despair (not shown in the graph). This way, the basic set of 8 emotions is complemented with 28 more complex variations, better representing the complexity of human emotions.

Cambria et al. [32] reinterprets the Plutchik model arranging primary emotions into four interrelated yet independent dimensions: pleasantness, attention, sensitivity, and aptitude. This model, known as the Hourglass Model, is depicted in Figure 2.1b. Cambria et al. [32] argue that the transition between different emotional states is not linear, thus they model this variation with a negative Gaussian function that gives the model its hourglass shape. Each of the 4 dimensions is divided into 6 levels of



(a) Wheel of Emotions (public domain figure).



(b) The Hourglass of Emotions. (Figure from Cambria et al. [32]).

Figure 2.1 – The Wheel of Emotions (Plutchik [148]) and the Hourglass of Emotions (Cambria et al. [32]).

intensity, producing 24 basic emotions. Similarly to Plutchik's model [148], the combination of fundamental emotions produces more complex emotions, for example, the combination of joy and trust produces love.

The advantage of using discrete emotions as a representation of emotional feeling is that it allows the use of common language to describe emotions, facilitating their understanding. However, one may argue that it may not be enough to use a limited set of words to represent an emotion, as discrete emotions may not cover the entire spectrum of feelings. Therefore, other ways of representation have been developed, which are presented below.

2.1.2.2 Continuous Representation: Dimensional Models

According to Ekman [57], proponents of the basic emotion theory maintain that "there are a number of separate emotions, that differ one from another in important ways". Ekman [57] also states that this perspective "is in contrast to those who treat emotions as fundamentally the same, differing

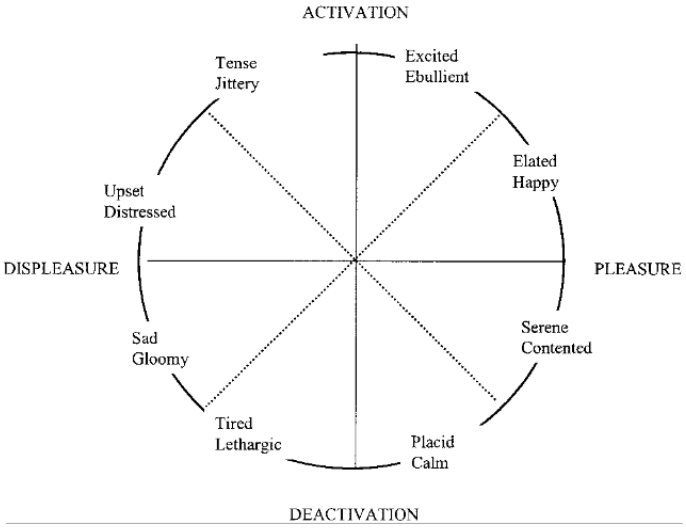


Figure 2.2 – Circumplex model of emotions, with affective concepts mapped on it. (Figure from Russell [157]).

only in terms of intensity or pleasantness". It is this second point of view, that emotions can be represented as continuous values that vary in terms of intensity or pleasantness, that is reviewed in this section.

Arguably, the most common model that follows this perspective is the circumplex model, developed by Rusell [155]. This model was developed under the premise that affective concepts can be defined in terms of two orthogonal dimensions: one dimension is pleasure/displeasure and the other is the degree of arousal. The circumplex model is depicted in Figure 2.2, with some affective concepts mapped on it. The pleasure/displeasure axis is commonly known as the valence axis. Thus, valence can be understood as how good/positive or bad/negative the feeling is. The arousal axis (also known as activation/deactivation axis) represents how intense the feeling is. Some authors, such as Russell and Mehrabian [159] or Mehrabian [133], add a third axis that represents dominance. The dominance axis indicates how in control a person feels in a situation, going from submissiveness to dominance. This 3-axis model is known as the PAD model, for Pleasure-Arousal-Dominance.

It may be argued that dimensional models are less intuitive and could be more difficult to interpret than using discrete emotions. Nevertheless, dimensional models, in particular the circumplex model, are probably the most commonly used models to measure emotional experiences [44].

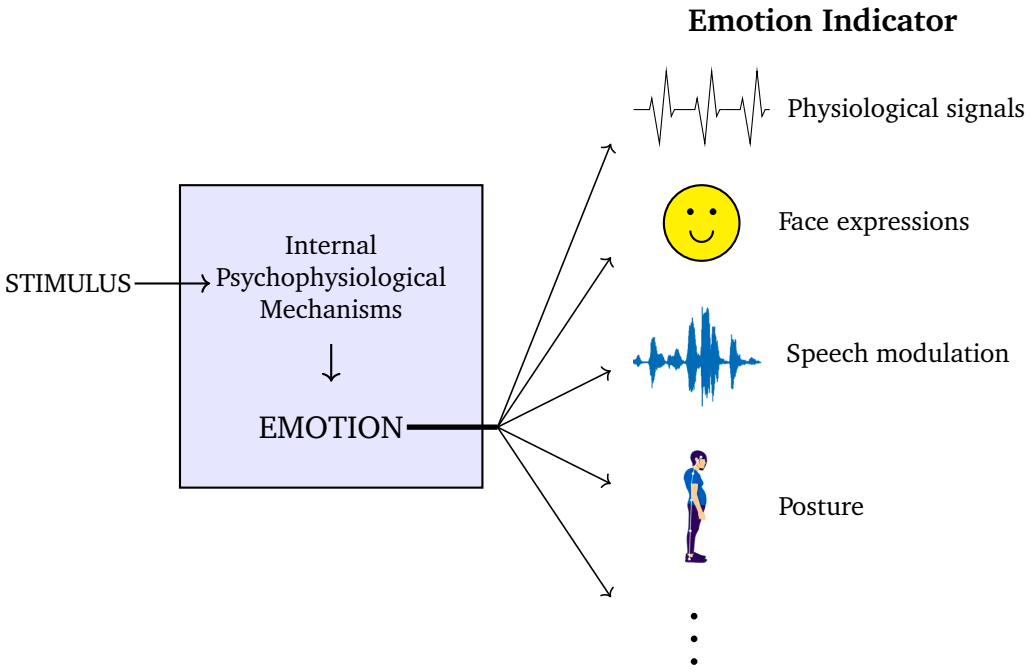


Figure 2.3 – A simplified perspective of emotion generation.

2.2 Emotion Recognition

Emotion recognition refers to identifying the emotion that a person is feeling using signs or expressions from this person. In this thesis, our goal is to design a system capable of performing emotion recognition automatically. For this, Section 2.2.1 provides a simplified perspective of emotions that will be helpful in defining emotion recognition. Then, Section 2.2.2, explains how an emotion recognition system could work, describing the different signals that could be used as inputs for such a system, and the outputs that it may produce.

2.2.1 Emotions, a simplified perspective

The human emotion mechanisms can be viewed from a simplified perspective. Although this view takes elements from the psychological research that was reviewed in previous sections, it is not a representation of the real mechanisms that are involved in emotions. This simplified perspective is presented in Figure 2.3, and models emotions as if they were produced by a simple input-output system. The input of the system is a stimulus (e.g. learning good or bad news, watching a scary scene in a movie, etc.) that

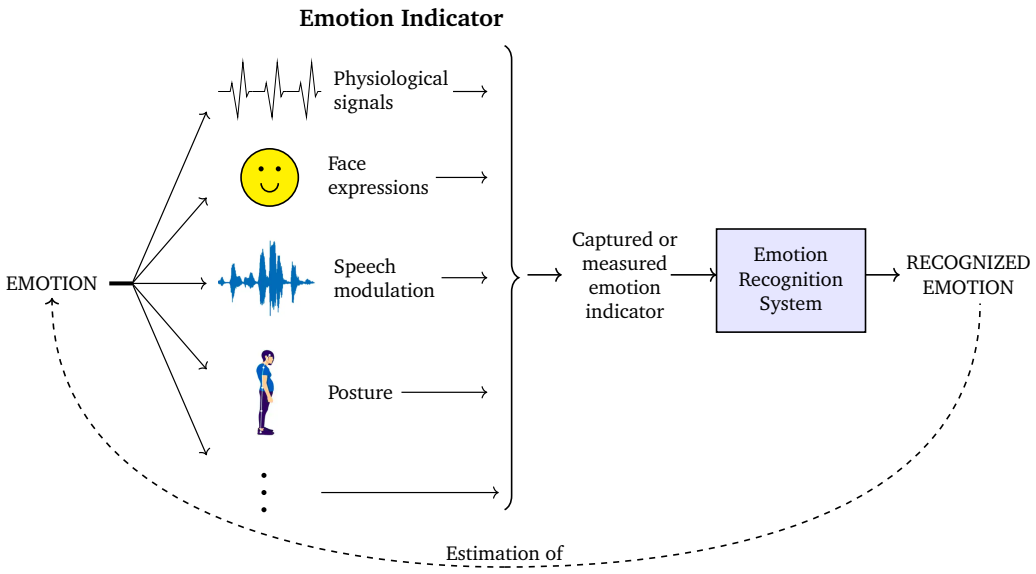


Figure 2.4 – A depiction of an emotion recognition system.

generates an emotion, and the output is an indicator of that emotion. In this context, we denote as an indicator of emotion any expression of emotion that can be observed, maybe with the help of an instrument, by an external observer. Some examples of these indicators are facial expressions, voice tone, and physiological signal features.

2.2.2 Emotion Recognition System

To explain how an emotion recognition system may work, we use the simplified perspective explained above. The idea is to instrument, i.e. measure or capture, the indicators of emotion produced during emotional events. For example, these measurements can be made with sensors that might exist as a part of a smart environment [150], or these measurements can be done for healthcare purposes. Then, these measurements can be used as inputs for an emotion recognition system, that gives as an output an estimation of the emotion that originally produced the indicator [110]. This process is depicted in Figure 2.4.

Note that in reality, the indicators of emotion used as inputs for the recognition system are not necessarily produced directly by emotions, since the interactions between emotions and indicators of emotions are more complex, as described in Section 2.1. This is why we use the simplified perspective from Section 2.2.1, because it allows us to see these indicators

Figure 2.5 – Facial expressions (by Icerko Lýdia/CC-BY-3.0).

as a consequence of emotions, without having to worry about the real interactions between them.

It can be noted from Figure 2.4 that to build an emotion recognition system it is necessary to identify the different expressions that can be used as indicators of emotion. In other words, it is required to define what can be used as input for an emotion recognition system, which is done in Section 2.2.3. In addition, a way to represent the recognized emotion is needed, meaning that it is necessary to define how the output of the system should look, which is done in Section 2.2.4.

2.2.3 Input of an Emotion Recognition System: Capturing Indicators of Emotion

There are several indicators of emotion that may be used as inputs for an emotion recognition system. Some examples of such indicators include facial expressions, speech, and body postures. There are also indicators of emotion in Electrocardiogram (ECG), Electroencephalogram (EEG), Electrodermal Activity (EDA), Electromyogram (EMG), Respiration (RESP), Temperature (TEMP), and other physiological signals. These indicators can be classified into four modalities: visual, audio, text, and physiological signals, which are reviewed below.

2.2.3.1 Visual Modality

Human emotions may be directly perceived from multiple channels using vision. For example, humans communicate emotional states with facial expressions and these can be directly observed and measured with computer vision. Figure 2.5 shows different facial expressions that display different emotions. In addition to facial expressions, there are other visual indicators that reflect the emotional state like gaits [24], body gestures [205] and posture [182].

Charles Darwin [48] described facial expressions as innate and universal, and he also theorized that these expressions evolved through interaction with the physical environment; for instance, he speculated that the display of disgust was initially linked to spitting spoiled food items [44]. Other authors (e.g. Ekman [56], Tomkins and McCarter [186], Larsen and Frederickson [117]) have also studied the relation between facial expressions and emotions, and how these expressions show the emotional state of a person. Many of these works have also suggested that facial expressions, and the interpretation of these expressions, are universal across cultures, although this has been heavily debated in the psychological research community, by critics like Russell [156]. Another characteristic of facial expressions, and visual indicators of emotion in general, is that they can be faked. For instance, talented actors are capable of displaying different emotions in a convincing way.

Ekman and Friesen [53] developed a widely used system, the Facial Action Coding System (FACS), that can be used to decode emotions from the activation patterns of groups of facial muscles, deconstructing facial expressions into distinct muscle movements called Action Unit (AU). For example, AU 4 is associated with lowering the brow. FACS use 28 main AUs, and provides additional ones to code the position and movements of the head and eyes. The activation strength of each AU can also be coded, using letters A (minimum) to E (maximum). A variation of this system was developed by Friesen and Ekman [70], that considers only AUs related to emotions.

From previous paragraphs, it is evident that it makes sense to use visual indicators as input for an emotion recognition system. Typically, this input is in the form of standard images, but authors like Wang et al. [198] have experimented with using thermal infrared images. Visual inputs may be used directly, that is, feeding into the system raw images or video frames. Another option is to code the visual information, for example, code the facial

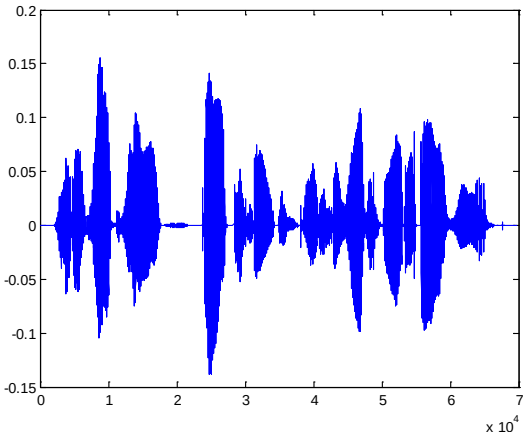


Figure 2.6 – An example of a speech signal. (Figure from Riadh et al. [152]).

expressions using FACS, and feed the coded information into the emotion recognition system. Even with the debate around its universality from critics like Russell [156], visual expressions remain a useful input for emotion recognition.

2.2.3.2 Audio Modality

As noted by several authors, like Kappas et al. [100], Bachorowski et al. [13], and Scherer et al. [165], expressions of emotion can be found in human speech. In fact, according to Scherer et al. [165], the effects that emotional responses have on respiration and muscle tension (especially in the larynx muscles) produce acoustical effects in speech like changes in its fundamental frequency, intensity, harmonic energy, and others. These authors also state that the arrangement of the vocal tract during certain emotional episodes will give preference to specific filter traits of the vocal tract, consequently impacting the distribution of energy in the spectrum. Moreover, they affirm that the changes in the larynx muscles will appear regardless if the subjects decide to speak or not.

In an effort to standardize the acoustic parameters used for speech analysis, Eyben et al. [64] proposed the Geneva Minimalistic Acoustic Parameter Set (GeMAPS). They choose parameters that are good indicators of changes in speech characteristics during emotional episodes, taking into account how frequently and successfully those parameters have been used in past literature. They also took into account the theoretical significance of the different parameters. Eyben et al. [64] presented a minimalistic parameter set, consisting of 18 low-level descriptors like pitch, loudness, jitter (deviations

from the fundamental frequency), and other attributes including spectral characteristics. They also introduced an extended parameter set, called extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), that added additional parameters to the minimalistic set with additional spectral and frequency-related attributes, bringing the total number of parameters to 88.

In summary, human speech can communicate emotions and therefore it can be used as input for an emotion recognition system. Besides speech, other useful clues could be found in audio signals, like laughter, crying, etc. Audio inputs can have the form of raw audio signals, like the speech signal depicted in Figure 2.6. Another option is to use acoustic parameters like the ones from eGeMAPS.

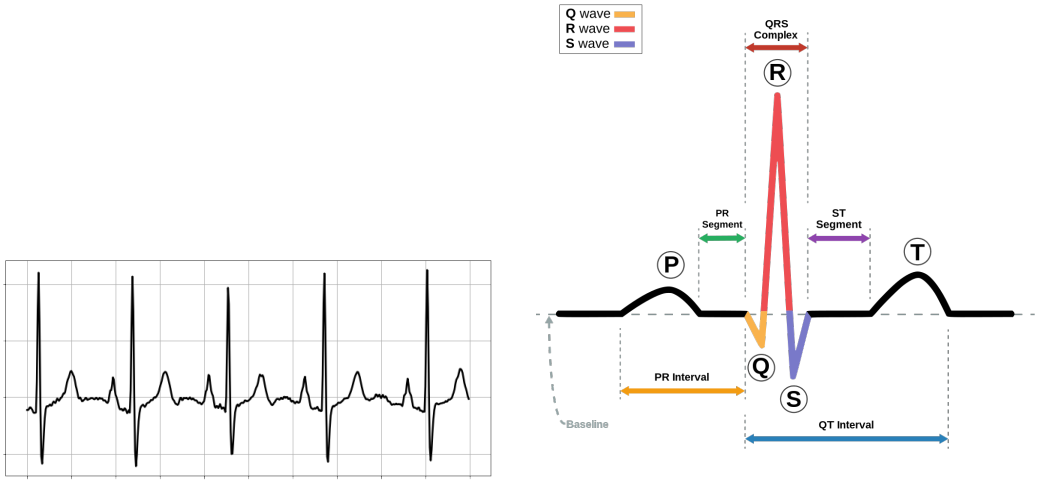
2.2.3.3 Text Modality

In addition to using audio signals or parameters from those signals as inputs for an emotion recognition system, it is also possible to extract the words that the subject is saying, i.e. extract transcripts, and use these words as input for the system [106, 167, 213]. These extracted words belong to a distinct modality, specifically the text modality, differing from audio due to their inherently different nature.

In general, affective analysis of text encompasses other tasks beyond recognizing emotions using transcripts; for example, written reviews of a user can be analyzed to determine if the attitude of the user is positive, neutral, or negative [137]. This type of analysis is commonly known as sentiment analysis, where the goal is to determine the valence of a piece of text (positive, neutral or negative), as mentioned by Mohammad [137].

2.2.3.4 Physiological Signals

According to Coppin and Sander [44], bodily reactions are considered an element of emotion by all prominent contemporary theories. This shows that there is a strong link between physiological reactions and emotions. For this reason, its usage as an input for an emotion recognition system is very appealing. In addition, physiological signals may convey information about emotions that is not externally expressed. Another advantage is that, different from visual and audio modalities, it is difficult to modulate these signals voluntarily, making it more difficult to *fake* an emotion. Below, some physiological signals that can be used as inputs for emotion recognition are reviewed.



(a) Example of an ECG signal.

(b) Elements of an ECG waveform.

Figure 2.7 – ECG signal and elements (Fig. (b) by A. Atkielski/Public Domain)

Cardiovascular Signals

There are several works that demonstrate the link between emotions and cardiovascular measurements like heart rate, heart rate variability, systolic and diastolic blood pressure, pre-injection period, and others. For example, Delplanque et al. [50] show that heart rate is a good indicator of valence. Also, Wu et al. [207] investigated the influence of emotions on cardiac responses, and found that certain emotions were reflected in changes in heart rate and heart rate variability.

Works like Delplanque et al. [50] and Wu et al. [207] show that there is a relation between cardiovascular activity and emotions. One way to capture cardiovascular information is to use ECG. ECG are recordings of electrical activity in the heart, typically collected by placing electrodes in the chest. Specifically, ECG register the changes in electrical potential difference during depolarization (contraction) and repolarization (relaxation) of the heart [200].

To better understand ECG signals, Figure 2.7a shows a sample of such a signal, whereas Figure 2.7b shows the elements that can be identified in a ECG waveform. From those figures, it can be observed that the elements of an ECG signal are: P wave, which is produced by atrial depolarization (i.e. atrial contraction); QRS complex, produced by ventricular depolarization (i.e. ventricular contraction); T wave produced by ventricular repolarization (i.e. ventricular relaxation). All these parts together are known as a PQRST

wave, and each heartbeat produces one of these waves.

Several parameters that can be useful for emotion recognition can be extracted from an ECG signal. For example, by identifying the QRS complexes present on a ECG signal, the time between two consecutive R peaks (called the RR interval) can be calculated, and from this quantity parameters such as heart rate and heart rate variability can be computed, which have been demonstrated to be related to emotions [50, 224]. Additionally, time-domain statistical parameters can be obtained, like the mean and median RR intervals, the standard deviation of RR intervals, and others. Furthermore, frequency-domain parameters like the total power in the full frequency range and the power in different frequency bands can also be extracted from ECG signals. All these parameters have been shown to be useful as inputs for an emotion recognition system [77, 90, 161].

Given that cardiovascular activity is linked to emotions, it makes sense to use ECG signals as input for an emotion recognition system, either in its raw form, or using different parameters extracted from those signals.

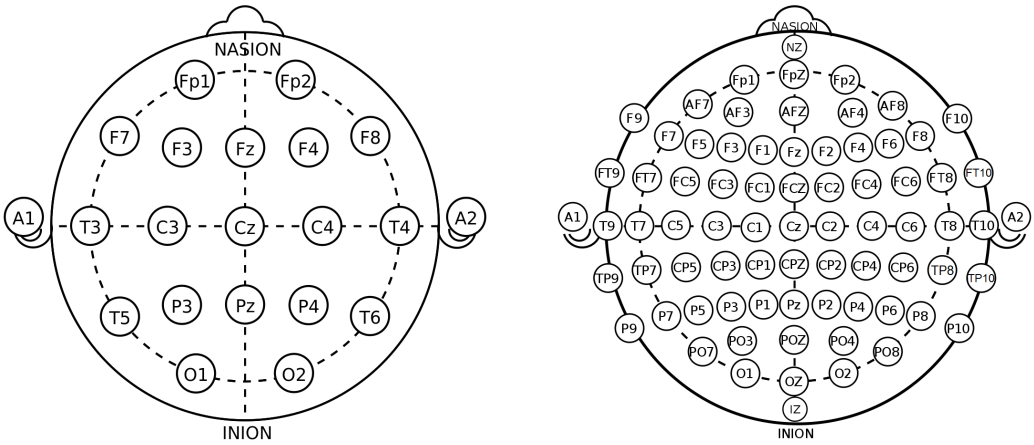
Brain Imaging

According to the appraisal theory, there is a cognitive evaluation of a stimulus before an emotion is elicited. Therefore, when an emotion is felt it should be reflected in the brain activity. For this reason, it is appealing to image the brain, i.e. measure brain activity, and use these measurements to perform emotion recognition.

There are several ways to monitor brain activity, for example through EEG, Functional Magnetic Resonance Imaging (fMRI), and Magnetoencephalography (MEG). Here, we give details about EEG, as it is one of the most accessible methods to monitor brain activity because it is cost-efficient, it can be acquired with portable systems, and it does not require intensive training to use.

EEG signals are acquired through electrodes placed on the scalp, which record the electrical activity of the brain. Typically, the electrodes are placed using the international 10-20 system or the 10-10 system, which are depicted in Figure 2.8. A more dense placement system exists, called the 10-5 system [142]. An example of EEG signals is depicted in Figure 2.9, where each of the showed waves is a signal acquired by one electrode.

From the appraisal theory of emotions, it can be argued that brain activity should be related to emotions. Therefore, it is feasible to use EEG signals as input for an emotion recognition system. Moreover, these types of signals



(a) 10-20 EEG electrode placement system (public domain figure).

(b) 10-10 EEG electrode placement (fig. by B. C. Oxley/CC0).

Figure 2.8 – EEG electrode placement sytems.

are more accessible compared to other options to monitor brain activity, therefore they are a good solution to incorporate brain imaging in such emotion recognition systems.

Respiration

It has been shown that variations in respiration occur during changes in emotion [183]. For example, Noguchi et al. [141] showed that fear and anxiety increase the respiration rate. Also, Boiten [25] found clear effects on inspiration, inspiratory pauses, tidal volume, and breathing irregularity, when subjects watched an emotionally loaded film. These works show that it is feasible to use respiration as input for an emotion recognition system.

Respiration may be measured by detecting the electrical impedance changes that occur in the chest during the respiration process. A good correlation exists between this impedance and the volume of breathed air. An example of a respiration signal is depicted in Figure 2.10.

Electrodermal Activity (EDA)

EDA is the measurement of changes in the electrical conductance of the skin produced by perspiration or sweat. Although sweat glands have thermoregulation as their primary function, glands in the hands and in the plantar regions have been shown to have a relation with behavior [89]. In fact, EDA has demonstrated to be an indicator of arousal [36].

As depicted in Figure 2.11, EDA signals are slow-changing, typically

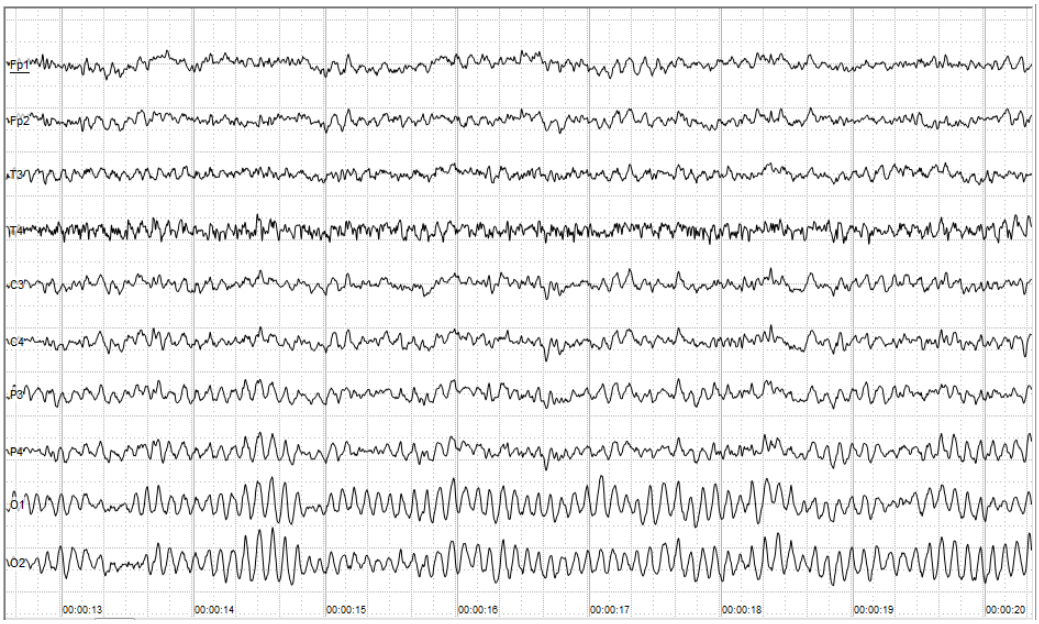


Figure 2.9 – EEG signals (by A. Cherninsky/CC-BY-SA-4.0).

characterized by rises and falls in the conductance values. Each of these "rises and falls" is known as an Electrodermic Response (EDR). As described in Boucsein [26], a EDR can be a response to a specific stimulus, like responses 1 and 6 in Figure 2.11, or could be spontaneous (maybe from an unknown stimulus), like responses 3 and 4, or can be undetermined, like responses 2 and 5.

Various parameters can be extracted from a EDA signal, like the amplitude of a EDR or the number of EDRs in a period of time. Analysis of these parameters has been widely used in psychological research [134], including detecting emotions [89]. Therefore, EDA, either raw signals or parameters extracted from those signals, should be useful as input for an emotion recognition system.

2.2.3.5 Discussion on Inputs for Emotion Recognition

In this section, we have described several indicators for emotional state that can be used as input for an emotion recognition system. Note that we used psychological literature to justify the reasons why these indicators should work as inputs for emotion recognition. We avoided basing the justification on contributions that have successfully developed an emotion recognition system using the indicators we reviewed. We did this because

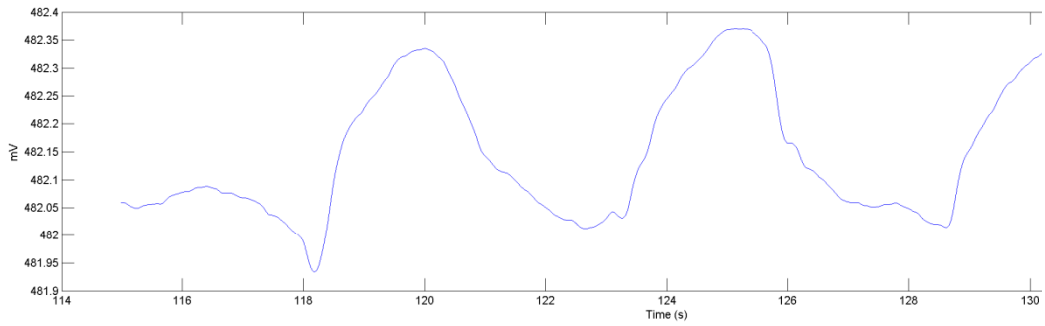


Figure 2.10 – Respiration signal. The measured millivolts (mV) are correlated with the volume of respired air. (Fig. from Shimmer3 User Manual [170])

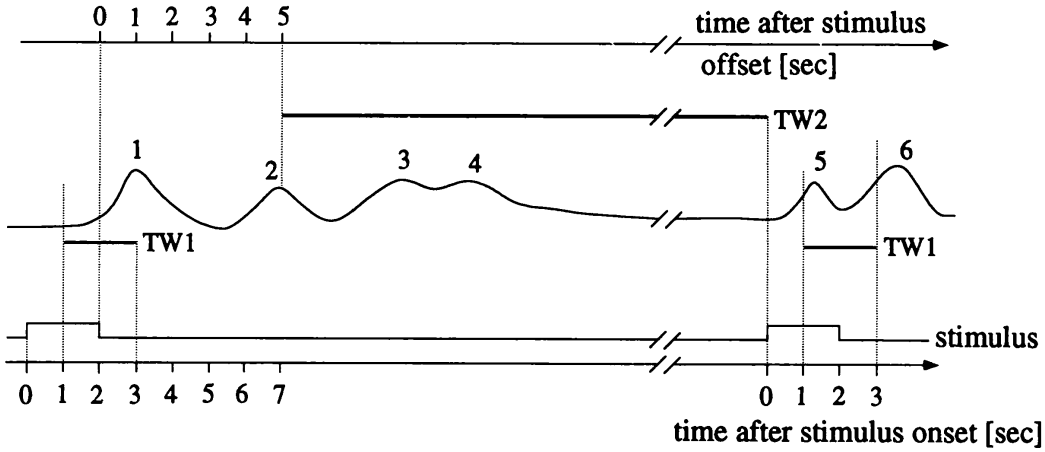


Figure 2.11 – Electrodermal Activity signal (Fig. by Boucsein [26]).

we believe it gives a stronger ground on the validity of those indicators as inputs for an emotion recognition system.

Nevertheless, there are plenty of works addressing emotion recognition that have used the indicators reviewed in this section. For example, Filntisis et al. [68] use visual inputs, Trigeorgis et al. [189] use audio inputs, Xiao et al. [210] use ECG signals, Hajlaoui et al. [82] use EEG signals, and Chatterjee et al. [37] use EDA. These works further validate the usefulness of those indicators as inputs for emotion recognition systems.

In the first part of this thesis, we use ECG and EEG signals. As our global objective falls within the framework of monitoring the mental health of frail people, these signals could already be acquired for other health purposes.

Moreover, these signals can already be acquired with portable, wireless, low-cost, and off-the-shelf equipment, as is done to build the AMIGOS dataset [135] and the DREAMER dataset [102], and thus we envision that the acquisition of these signals will become even less intrusive and easier to perform in the future. Another advantage is that datasets that contain these signals are abundant, if we include datasets that have not necessarily been collected for affective tasks, but with other goals like medical research. This is important for performing the self-supervised pre-training technique that is introduced in Chapter 3. In addition, in Chapter 4 we use both ECG and EEG at the same time, hypothesizing that doing this will improve the results when doing emotion recognition. We believe this should be the case since those signals may carry complementary information since they come from different mechanisms inside the body: EEG signals monitor brain activity, thus they are more related to cognition, while the characteristics of EEG are related to the autonomic nervous system, thus are the result of a more primary reaction.

In later chapters, we consider the use of auditory and visual modalities in combination with physiological signals, performing emotion recognition in a multimodal fashion. Using similar reasoning for combining EEG and ECG signals, we believe that the information carried by audio, visual, and physiological signals may be complementary. In fact, in physiological signals, there might be information about the emotion being felt that might not be externally expressed. There is an additional advantage of using multiple types of inputs: the system could be built to be robust to missing modalities. For this, using heterogeneous sources (e.g. audio, visual and physiological signals), improves the usability of a system in varied situations, having different sensing capabilities which might not all be available at all times.

2.2.4 Output of an Emotion Recognition System: Emotion Representation

When designing an emotion recognition system, it is necessary to define what will be its output. In other words, it is necessary to define how the recognized emotion will be represented by the system. From the review in Section 2.1.2, it can be concluded that it is possible to use as output a discrete or a continuous representation of emotions.

A system that identifies an emotion using discrete representations will have as output the name of the identified emotion. For example, the output could be one of the following emotions: happiness, fear, disgust, anger,

sadness, or surprise. On the other hand, a system that uses a continuous representation may output numerical values of arousal and valence.

In this thesis, we decide to represent emotions using arousal and valence. We do this because we are mainly interested in determining how a person is doing emotionally, so we want to know the pleasantness of the emotion being felt (valence) and the intensity of that emotion (arousal). We believe that for this objective of knowing how the person is doing emotionally using arousal and valence is more convenient than using discrete emotions. In addition, we decide to use only arousal and valence and no other dimensions of emotion because we are more interested in elementary affective feelings. In this sense, we adopt a point of view similar to the one described by Rusell and Barrett [158], who consider other dimensions (e.g. dominance) more related to the event that produces the emotion, and thus outside of being part of the elementary feeling.

More specifically, the first part of this thesis uses high and low categories of arousal and valence as outputs for an emotion recognition system. Whereas, for the second part of this work, the output of the system is a numerical value of arousal and valence. When using high and low categories, since arousal and valence are continuous values, a threshold value can be used to determine those high and low categories. This threshold value could be, for example, the average arousal/valence value of the considered samples used to develop the system.

2.2.5 Putting it all Together

We now have all the elements necessary to build an emotion recognition system. We can also provide a definition of what emotion recognition is in the context of this thesis:

Definition 2.1. ***Emotion Recognition:*** *Estimation of the emotional state of an individual using one or more observed indicators (visual, auditory, physiological or other perceptual channels).*

An depiction of such a system is shown in Figure 2.12. Note that, as discussed in this section, the inputs can be either raw signals or parameters extracted from those signals. Moreover, the system could draw information from only one modality (i.e. a single-modality system) or several modalities (i.e. a multimodal system). As discussed before, the output of the system can be discrete emotions (e.g. happiness, fear, disgust, etc.) or, as is the case for this thesis, numerical values or categories of arousal and valence.

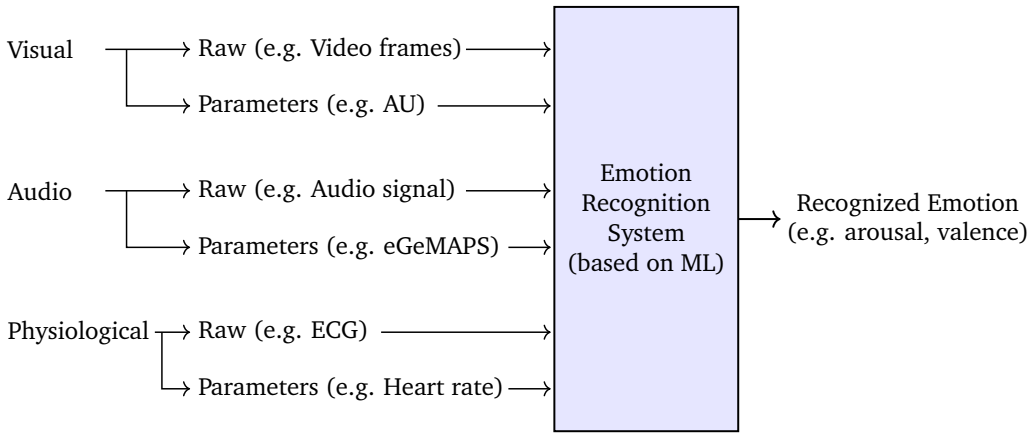


Figure 2.12 – An emotion recognition system: putting it all together

At the heart of an emotion recognition system, there is a model capable of describing the relation between the different indicators of emotion and the emotion behind those indicators. In other words, this model is the mechanism that receives as inputs indicators of emotion and produces as output a representation of the emotion that produced those indicators. When the emotion representation is a category, as it is in the first part of this thesis where high and low categories of arousal and values are used, the model has to perform a classification task. While if the emotion representation is a numerical value, as in the second part of this thesis where continuous values of arousal and valence are used, the model performs a regression task. A model of such characteristics can be implemented using Machine Learning (ML).

Through this thesis, we design single architectures to recognize arousal and valence and train the same architecture one time to recognize arousal and a second time to recognize valence. Instead of this approach, other options are designing a multitask system, that recognizes arousal and valence at the same time; or designing an architecture for arousal and a different one for valence. We prefer having unified architecture trained independently for arousal and valence, as it is a good trade-off between having to design two different specialized systems, and having a single global system that might sacrifice individual performance.

In real-life scenarios, there might be additional challenges that need to be addressed when designing a system for emotion recognition. One challenge has to do with noisy and incomplete inputs; for instance, audio signals might incorporate background noise, or the face of a person might

be partially occluded. In our work, this challenge is partially addressed in Chapter 6, where we assume that an input that is too noisy to be useful can be detected as such, so it can be discarded, and the system should keep working with the remaining inputs.

Another challenge is that an emotion recognition system has to deal with the diversity in individual reactions to stimuli. In fact, as stated by Hajlaoui et al. [81], across different subjects the same stimulus may produce different emotions, and the same emotion may raise different physiological responses. In addition to this, there might be cultural differences in emotional responses [119, 113]. Thus, personalized systems may be envisaged, where data from a subject can be used to train a personal model to recognize emotions from that same subject, using for example few-shot learning techniques if data from this person are scarce. Different from this, in this thesis, we are more interested in developing a general system, that works at least for people within the same demographics. The reason for this is that we envision the emotion recognition system as a part of a smart environment where the emotional state of frail people can be monitored. Therefore, it is important that the deployment of the emotion recognition system is easy. If a personalization step is needed, users might need to collect and label their own data, which could be an obstacle to the use of the system.

2.2.6 Models for Emotion Recognition

An emotion recognition model can be implemented using classical ML techniques, like Gaussian Naive Bayes, Support Vector Machine (SVM), k-Nearest Neighbors, and others. Several works have used this technique for emotion recognition, using different inputs like speech [115], physiological signals [90], and faces [76]. Typically, these types of models use different parameters extracted from the input signals. Generally, these parameters are chosen using domain expertise, and for this reason, the chosen parameters are called *engineered features*.

An alternative to classical ML techniques is the use of Deep Learning (DL) models. These models typically exhibit superior performance to classical ML approaches. In fact, Abbaschian et al. [2] show that in speech emotion recognition, for some tasks, the accuracy increases between 70% and 90% when using DL compared to classical ML. In addition, Maithri et al. [131] reviewed several works that use classical ML and DL techniques for emotion recognition using EEG, facial, speech, and multimodal signals, and conclude that using DL leads to higher performance. An additional advantage of DL

models is that they can process raw signals instead of engineered features, being capable of learning intermediate representations better aligned to the addressed problem, and therefore leading to improved performance [192]. When DL is employed, Convolutional Neural Network (CNN) may be used to process images [149], audio [2], and physiological signals [161]. In addition, to model sequential information in the inputs Recurrent Neural Network (RNN) and its variations (Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), etc.), may be employed [6, 192].

Recently, the Transformer [196] has emerged as successful DL architecture in fields like Natural Language Processing (NLP) [51], computer vision [83], and speech processing [118]. It has been also used successfully for emotion recognition [95, 190, 208]. An advantage of the Transformer is that it is capable of effectively capturing long-range dependencies in the input sequence, in contrast with RNNs that may struggle to capture such dependencies. Moreover, it allows parallel processing of the sequence, as the output does not depend on past states. These benefits are the result of the Transformer relying entirely on attention mechanisms to build representations of the input, without using recurrent or convolutional networks. Moreover, the attention mechanisms of the Transformer are capable of generating representations that pay more attention, i.e. give more weight, to the relevant parts of the input. For this reason, we use a Transformer as the backbone of the different architectures that we introduce in this thesis.

2.3 Overview of the Transformer

A Transformer [196] combines multiple layers of encoders and decoders to provide an extremely powerful architecture for processing signals. While the Transformer was originally designed for NLP tasks, such as machine translation, it is suitable for use in other tasks requiring interpretation of other types of sequences, like frames from a video [83], sound signals [118], and physiological signals [223]. Thus, it is feasible to use a Transformer to estimate arousal and valence from the different modalities discussed in Section 2.2.3. Importantly, the transformer is well-suited for interpreting multiple signal modalities.

The remainder of this section describes in detail the Transformer architecture, mainly summarizing the original paper by Vaswani et al. [196].

2.3.1 Attention

A key innovation of the Transformer is the use of attention to simplify processes for encoding and decoding information. The attention function of the Transformer can be characterized as a mapping of a query to a set of key-value pairs, where all of them are vectors. The output of this function is a weighted sum of the values, where the weight for each value corresponds to the compatibility of the query with the corresponding key.

2.3.1.1 Scaled Dot-Product Attention

The Transformer employs an information selection technique referred to as *scaled dot-product attention* to associate related parts of the input sequence. The Transformer expresses information using a system of symbolic tokens. Tokens are encoded in a key-value representation that enables tokens to be easily associated with related parts of a signal expressed as queries. Keys and queries are implemented using a trained linear encoding that can be used to retrieve tokens that are related to a query by a simple matrix multiplication.

If queries come from sequence $\hat{Q}$, keys from sequence $\hat{K}$, and values from sequence $\hat{V}$, query (Q), key (K) and value (V) matrices are obtained as follows:

$$Q = \hat{Q}W^Q \qquad K = \hat{K}W^K \qquad V = \hat{V}W^V, \tag{2.1}$$

where $W^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, and $W^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are learnable projection matrices. Also, d_{model} corresponds to the dimensions of the intermediate representations inside the Transformer, as well as its output dimension, while d_k and d_v are the dimensions of the key and value vectors, respectively.

Scaled Dot-Product Attention is computed by performing the dot product between a query and all the keys, dividing each result by $\sqrt{d_k}$, and then using the softmax function to obtain the weights for each value. With all the queries, keys and values packed in matrices Q, K and V, this can be done simultaneously for all the queries using the following expression:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2.2}$$

2.3.1.2 Multi-Head Attention

The authors of the Transformer architecture found it beneficial to obtain h versions of queries, keys and values with different learnable projection matrices W_i^Q, W_i^K, and W_i^V:

$$Q_i = \hat{Q}W_i^Q \qquad K_i = \hat{K}W_i^K \qquad V_i = \hat{V}W_i^V. \tag{2.3}$$

Then, the attention function defined in Expression 2.2 is computed for each version of the queries, keys and values, obtaining the output of a single head:

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i). \tag{2.4}$$

The outputs of each head are concatenated and projected again using the learnable matrix $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$. Thus, Multi-Head Attention (MHA) is defined as:

$$\text{MHA}(\hat{Q}, \hat{K}, \hat{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O. \tag{2.5}$$

The dimensions d_k and d_v are chosen as $d_k = d_v = d_{\text{model}}/h$. With this, the computational cost is comparable to using full-dimension single-head attention.

Using multiple attention heads permits the creation of various representation subspaces, allowing the model to simultaneously attend to multiple information contexts at different positions of the input sequence.

2.3.1.3 Self-Attention and Cross-Attention

The input sequences $\hat{Q}$, $\hat{K}$ and $\hat{V}$ of the MHA module may be chosen to be the same sequence. In this case, each position of the sequence attends to the other positions from the same sequence. Accordingly, this case is called *self-attention*.

Another option is having a sequence as the query sequence $\hat{Q}$, and a different sequence as the key and value sequences $\hat{K}$ and $\hat{V}$. If this is the case, the first sequence attends each position of the second sequence, Note that in this case, both sequences do not need to have the same length. Having the query from one sequence and the key-value from a second sequence is known as *cross-attention*. Cross-attention can be used to associate information from different modalities, incorporating information from one modality, the source modality, into another modality, the target modality. In this

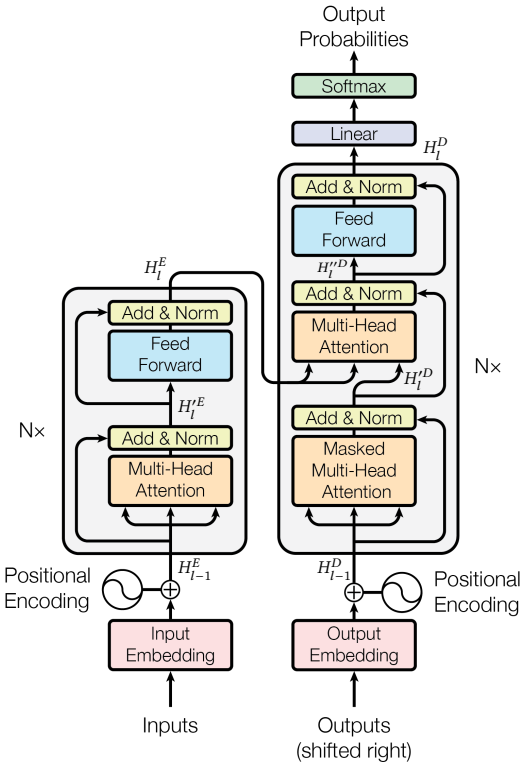


Figure 2.13 – Model Architecture of the Transformer (Fig. by Vaswani et al. [196]).

case, the keys and values come from the source modality, and the query comes from the target modality. This way, the target modality attends the source modality, effectively associating relevant information from the source modality, and incorporating this information into the target modality.

2.3.2 Transformer Architecture

The Transformer follows an encoder-decoder structure, using stacks of MHA and point-wise feed-forward Fully-Connected Network (FCN), as depicted in Figure 2.13. A description of the different components of the Transformer architecture is provided below.

2.3.2.1 Encoder

The encoder of a Transformer is composed of a stack of identical layers, where each layer is composed of a MHA module followed by a point-wise feed-forward FCN. The MHA performs self-attention, that is, all the queries, keys and values come from the same sequence, in this case, the input

sequence or the output of the previous encoder layer. Residual connections and layer normalization [11] is employed, as shown by the following expressions, that describe the process of a single layer of the encoder:

$$H_l'^E = \text{LayerNorm}(H_{l-1}^E + \text{MHA}(H_{l-1}^E, H_{l-1}^E, H_{l-1}^E)) \tag{2.6}$$

$$H_l^E = \text{LayerNorm}(H_l'^E + \text{FCN}(H_l'^E)), \tag{2.7}$$

where H_l^E is the output of encoder layer l, and H_0^E is the input sequence.

2.3.2.2 Decoder

The decoder is composed of a stack of Transformer Decoder Layers (TDLs). Each TDL is similar to an encoder layer, with the difference that an additional attention module is added to attend to the outputs of the encoder, i.e. perform cross-attention. In this cross-attention module, the queries come from the previous decoder layer, while the key and values come from the output of the last layer of the encoder. Thus, the expressions that define a single decoder layer are the following:

$$H_l'^D = \text{LayerNorm}(H_{l-1}^D + \text{MHA}(H_{l-1}^D, H_{l-1}^D, H_{l-1}^D)) \tag{2.8}$$

$$H_l''^D = \text{LayerNorm}(H_l'^D + \text{MHA}(H_l'^D, H_N^E, H_N^E)) \tag{2.9}$$

$$H_l^D = \text{LayerNorm}(H_l''^D + \text{FCN}(H_l''^D)), \tag{2.10}$$

where H_N^E is the output of the last encoder layer, and H_0^D is the input sequence of the decoder, which corresponds to the decoder outputs generated up to the current time.

2.3.2.3 Positional Encoding

The operations in the Transformer are permutation-invariant with respect to the order of the elements in the input sequence. Therefore, there is the need to inject information about the position that each element occupies in the input sequence. For this reason, *positional encodings* are added to the elements of the input sequence before they are fed to the encoder and decoder stacks. These encodings are designed such that a unique representation is generated for each position in the sequence. These position representations are vectors of size d_{model}, the same size as the input vectors, so the positional encodings and the vectors of the input sequence can be summed. Through the combination of these positional encodings with the inputs, the Transformer can take into account both the content and the order of inputs when processing those inputs.

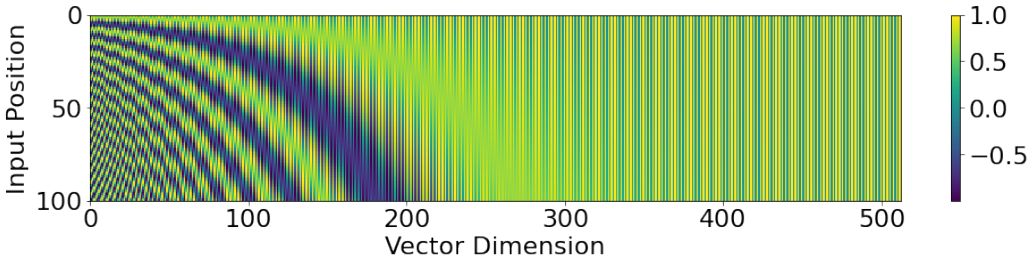


Figure 2.14 – Sinusoidal Positional Encodings.

In the original Transformer paper, fixed positional-encoding vectors are used. They are built using the following expressions:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{\text{model}}}) \tag{2.11}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}}), \tag{2.12}$$

where pos is the position to which the position encoding PE corresponds, and i is the dimension in the vector. Using sine and cosine functions produces a pattern of values that depends on the index inside each vector, but more importantly, depends on the position, as depicted in Figure 2.14.

An alternative to fixed positional encodings is to use learned positional encodings. The idea is to learn those encodings during training, at the same time as the rest of the parameters of the architecture. Then, if the input sequence is $X \in \mathbb{R}^{T \times d_{\text{model}}}$, the input with the positional information becomes $X + W_{PE}$, where the elements of $W_{PE} \in \mathbb{R}^{T \times d_{\text{model}}}$ are parameters that are learned during training.

2.3.3 Discussion About Transformers

When processing sequences, some favorable characteristics of the Transformer include its capacity to model long-range dependencies, and its ability to process the input sequence in parallel. In addition, the attention mechanism of the Transformer creates representations that are the weighted sum of the different elements of the input sequence. In a way, this weight is an indication of the importance of each element. This is visualized in Figure 2.15, which shows attention maps from different layers of a Transformed trained for image classification. In our case, if we are processing physiological signals for example, we can imagine that some parts of the signal may be more important than others, and thus is appealing to use an architecture capable of capturing this fact. Moreover, if multiple modalities

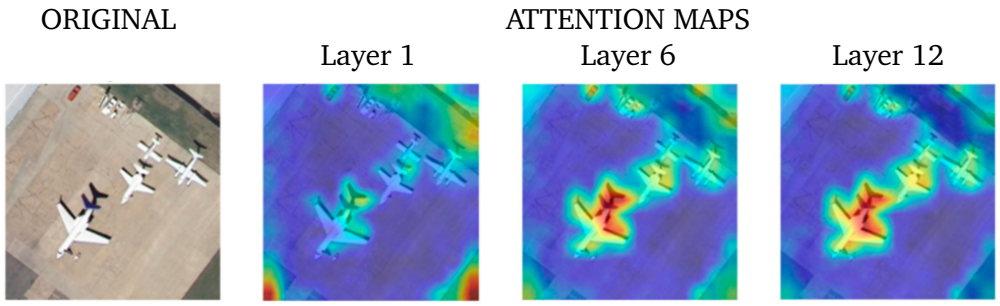


Figure 2.15 – Transformer attention maps at different layers. (Image from Bazi et al. [21]).

are used, the attention mechanism of the Transformer can be used to weigh the importance of each modality when building a representation that is the aggregation of all modalities, while also using self-attention to select and incorporate the most relevant information within each modality into that representation.

For the reasons stated above, in this thesis, we explore the use of the Transformer as the backbone of different architectures to perform emotion recognition, as we hope the characteristics of the Transformer will be beneficial for this task. To train the different models, datasets that contain labels of emotion are required. The next section briefly discusses some examples of such datasets. More details about the datasets used in this thesis can be found in Appendix A.

2.4 Datasets

There are several datasets that can be used to train models for emotion recognition, varying in the modalities that they provide and the labeling that they use. Regarding datasets that include physiological signals, which are the signals used in the first part of this thesis, there are datasets like DEAP [109], MAHNOB-HCI [177], and DECAF [1] that used costly, non-portable and non-wearable sensors to acquire physiological signals. Contrary to this, the AMIGOS [135], DREAMER [102], and ASCERTAIN [184] datasets acquire ECG and EEG signals using wireless, wearable, and off-the-shelf equipment.

To develop our approaches in the first part of the thesis, where ECG and EEG signals are employed, we use the AMIGOS and DREAMER datasets. In

addition to providing the physiological signals we are interested in, a reason to use AMIGOS and DREAMER datasets is that both datasets use portable, wearable, wireless, low-cost, and off-the-shelf equipment to acquire the signals. We believe that in the future the technology to acquire those signals could become even less invasive, thus capable of becoming part of a smart environment, where the emotional well-being of a frail person can be monitored with non-invasive wearable sensors. For these reasons, it is interesting for us to test our approaches with signals that are obtained with the equipment used in AMIGOS and DREAMER. More details about these two datasets are provided in Appendix A.

For the second part of this work, datasets that include multimodal data and time-continuous annotations of valence and arousal are required. Some datasets in this category are the Aff-Wild [214] and Aff-Wild2 [111] datasets, which were collected by continuously annotating emotional labels in YouTube videos. Another dataset is the SEWA DB dataset [112], which compromises audio-visual data of subjects recorded watching adverts and then discussing these adverts in a video chat. Also in this category of datasets, there is the RECOLA dataset [153], where participants were recorded during a video conference while completing a collaborative task. A final example is the Ulm-Trier Social Stress Test (ULM-TSST) dataset [179, 180], a multimodal dataset that was collected while inducing stress on the participants.

Given that our interest is to monitor the emotional well-being of frail people, it is interesting to use a dataset in which subjects go through a stressful situation, because it is important to be able to recognize the emotions of people going through difficult events. For this reason, in our work, we use the ULM-TSST dataset. Moreover, although we are not interested in measuring stress directly, it is an important factor to take into account for preserving the mental health of frail people. In fact, stress in old age may lead to serious health issues. For example, a study showed that vulnerability to stress in seniors is associated with an increased risk of Alzheimer's disease [202]. Therefore, it is interesting in the general framework where this thesis is developed to measure emotions elicited during stressful situations. More details about the ULM-TSST dataset is provided in Appendix A.3.

Through this chapter, we provide various definitions and assumptions used in this thesis, when developing emotion recognition systems. Also, the datasets that are used to train our models have been identified. The rest of this thesis provides details about our different contributions to the task of emotion recognition.

CHAPTER 3

EMOTION RECOGNITION FROM PHYSIOLOGICAL SIGNALS

In order to monitor the mental well-being of a frail person in a smart environment, it is desirable to have a system able to recognize emotions from the sensors present there. In this type of environments, wearable sensors can be used to collect physiological data, and thus it is interesting to predict emotions from these data. Since we envision emotion recognition systems as a part of a global health system, an advantage of using physiological signals is that those signals could already be acquired for other health purposes. Moreover, nowadays there are portable, wireless, low-cost, and off-the-shelf equipment capable of gathering some of these types of signals, therefore there is a potential that the acquisition of these signals become even less intrusive in the future.

This chapter describes a method to recognize emotions from physiological signals. It starts by providing a problem definition in Section 3.1, along with the challenges around solving this problem, followed by a review of current methods for emotion recognition from physiological signals in

Section 3.2. Next, our approach, which is the first contribution of this thesis, is detailed in Section 3.3, presenting a pre-trained Transformer-based model designed to perform emotion recognition from physiological signals. Finally, the results of evaluating our ideas are presented in Section 3.4.

3.1 Problem Definition and Challenges

3.1.1 Problem Definition

Using the definition of the emotion recognition problem provided in Section 2.2, we define the specific problem addressed in this chapter: How to recognize high and low categories of arousal and valence from raw physiological signals, or more specifically from raw Electrocardiogram (ECG) and Electroencephalogram (EEG) signals. We chose to use ECG and EEG signals since they can be acquired with portable, low-cost equipment, as done by Katsigiannis et al. [102] and Miranda-Correa et al. [135]. In addition, there is abundant unlabeled data of this type, which is important for our approach, as we shall see. Additional details about different aspects of the problem addressed in this chapter, including our motivations on why the problem is specified that way, are provided below.

3.1.1.1 Recognizing High/Low categories of Arousal and Valence

We formulate the emotion recognition problem as a classification problem, recognizing high and low categories of arousal and valence. We consider that setting the problem this way is appropriate for the general goal of this thesis: monitoring the emotional state of a frail person. Specifically, to accomplish this general goal, it is necessary to know if the monitored person is feeling a positive or negative emotion (high or low valence) and if the intensity of this emotion is high or low (high or low arousal). Moreover, other works that address the task of emotion recognition from physiological signals also aim to recognize high and low categories of arousal and valence [162, 161, 212], further validating our choice.

Nevertheless, it could be argued that a system that recognizes numerical values of arousal and valence could be preferable since it could provide a better understanding of the emotional state. We thus address this task in Chapter 5.

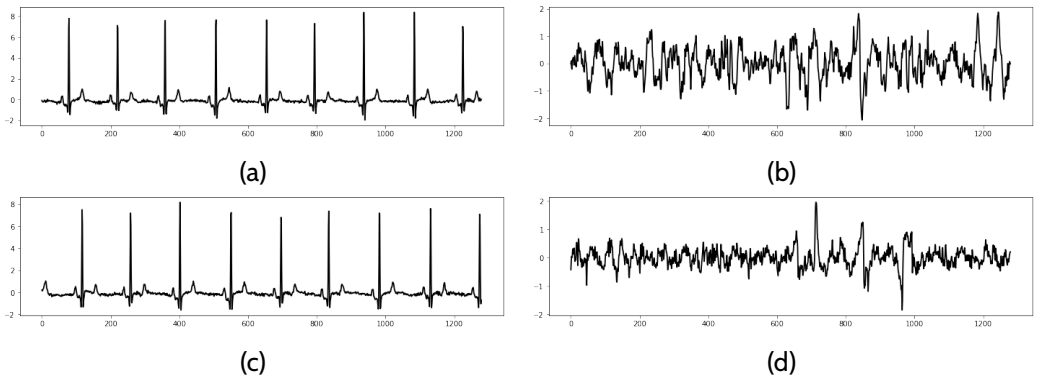


Figure 3.1 – Examples of raw physiological signals: (a) and (c) Electrocardiogram (ECG) signals, (b) and (d) Electroencephalogram (EEG) signals. The top row depicts signals labeled as high arousal and low valence, and the bottom row depicts signals labeled as low arousal and high valence.

3.1.1.2 Using Raw Physiological Signals

It has been demonstrated that physiological signals can be used to recognize emotions [171], and several works have emerged in this area using EEG [160, 218] and ECG signals [161, 162].

As discussed in Section 2.2.3, an emotion recognition system may use as input raw signals or parameters extracted from those signals. We will refer to the signals returned by sensors, including filtered or normalized versions, as *raw* signals. From these raw signals, different parameters may be extracted. Some examples of parameters that can be extracted from physiological signals are time and frequency-domain characteristics of the signal, like spectral entropy of the signal [168] or the spectral power of the signal [161].

A Deep Learning (DL) model may be designed to work with raw signals or with parameters extracted from the signal. However, when using parameters of the signals instead of raw signals, it is necessary to choose which parameters to use, which may require certain expertise, and choosing the most discriminant parameters may be difficult to do [173]. On the other hand, using raw signals does not require this expertise, and a well-designed deep neural network should be able to extract suitable features from the raw signal that can lead to good performance. Therefore, as we use a DL approach, we use raw physiological signals as input for our model.

Figure 3.1 shows examples of the two types of raw physiological signals used in this chapter. Figures 3.1a and 3.1c show examples of raw ECG

signals, and Figures 3.1d and 3.1d show examples of raw EEG signals.

3.1.2 Challenges

We identify two main challenges for our approach. The first challenge is the lack of sufficient labeled data to effectively train a deep neural network. The second challenge is how to process raw signals effectively.

Regarding the first challenge, a characteristic of DL approaches is that a satisfactory performance of these types of models typically depends on having enough labeled data to train the model. In fact, Sun et al. [185] claim that one of the reasons for the success of DL is the availability of large-scale labeled data. However, collecting physiological data and labeling it with labels of emotion is a long and expensive process, and therefore these types of datasets tend to be small. For example, the DREAMER dataset [102], which includes physiological signals with labels of emotion, contains around 23 hours of data. Meanwhile, for other tasks like action recognition from videos, there are datasets like the Kinetics-700 [35] with over 1800 hours of video. Not having enough data to train the model could lead to overfitting. Therefore, it is necessary to have a strategy that allows an emotion recognition model to perform adequately under these conditions. This challenge can be summarized as follows:

Challenge 3.1. *How to effectively train a DL model to perform emotion recognition from physiological signals, given that datasets of physiological signals with labels of emotion may not have enough data to do so.*

The second challenge is how to effectively process raw signals. Thus, this challenge can be stated as follows:

Challenge 3.2. *How to design a model that can extract from raw signals features suitable for emotion recognition, effectively modeling the different dependencies of the signals.*

A description of how we address these challenges is provided in Section 3.3. For the moment, in the next section, we review current methods of emotion recognition from physiological signals.

3.2 State of the Art on Emotion Recognition from Physiological Signals

This section provides a review of the literature and techniques relevant to the task of emotion recognition. This review is focused on the works that are pertinent to the problem and challenges described in Section 3.1.2. That is, we are mostly interested in works that deal with emotion recognition from ECG and EEG signals.

The discussion starts by reviewing works that use classical Machine Learning (ML) techniques in Section 3.2.1. Then, contributions that employ DL methods are covered in Section 3.2.2. After that, Transformer-based approaches are reviewed in Section 3.2.3. Then, an overview of training DL models with limited data is provided in Section 3.2.4. Finally, pre-training approaches, which are approaches that first pre-train the model with a pretext task using unlabeled data, are discussed in Section 3.2.5.

3.2.1 Classical Machine Learning Techniques for Recognition of Emotions

Classical ML techniques (non-DL techniques) employed for emotion recognition include Gaussian Naive Bayes, Support Vector Machines (SVMs), k-Nearest Neighbors, and Random Forests, among others. Generally, the works that employ these approaches use parameters extracted from the signals, instead of using raw signals. Since these parameters are the input features for the models, instead of parameters, we prefer to call them features. Typically, designing and selecting these features requires certain expertise, and usually these features are called *engineered features*.

Several authors have addressed the task of emotion recognition from physiological signals using classical ML techniques. Gjoreski et al. [77] use engineered features extracted from ECG and Electrodermal Activity (EDA) signals, and then process those features using several ML models like Random Forests, SVMs, Gaussian Naive Bayes, among others. Hsu et al. [90] work with ECG signals, extracting engineered features from those signals and using an algorithmic selection step to reduce the number of features to be used, in order to reduce computational complexity. Then, the selected features are used as input of a SVM. Shu et al. [172] use heart rate data as input. From that input, they extract engineered features like the mean difference in heart rates, heart rate range, heart rate mean

and variance, among others, and then they use an algorithm to select the most relevant features for the task. They trained five types of classifiers: k-Nearest Neighbors, Random Forest, Decision Trees, Gradient Boosting Decision Trees, and Adaptive Boosting. Subramanian et al. [184] extract several engineered features like heart rate statistics from ECG signals, and various statistical parameters from EDA and EEG signals. Then, they use those features as inputs for a SVM and a Naive Bayes model.

To better illustrate the kind of approach described in this subsection, the work of Hsu et al. [90] is detailed in Section 3.2.1.1, and the work of Subramanian et al. [184] is detailed in Section 3.2.1.2.

3.2.1.1 Emotion Recognition From ECG Signals

In the work of Hsu et al. [90], the goal is to recognize discrete emotions from ECG signals. To process the signals, they first obtain the RR intervals, i.e. the time between two consecutive R peaks. An R peak corresponds to the higher peaks typically seen on ECG signals (see Section 2.2.3.4 and Figure 2.7). Then, they get several features from the extracted RR intervals, as explained below.

First, they perform time-domain analysis to obtain 12 features. Some of the obtained time-domain features include the standard deviation of RR intervals, the root mean square of differences between adjacent RR intervals, and the number of successive RR intervals that differ more than 50ms, among others. Next, they perform frequency-domain analysis to obtain 13 features. Some of the frequency-domain features that they extract are the total power in the full frequency range, power in different frequency bands, frequency of the highest peak in different frequency bands, etc. Finally, they perform nonlinear analysis to obtain 9 additional features. Some examples of these additional features are the approximate entropy and the sample entropy. In total, 34 features are extracted.

Once the features have been extracted, they perform a feature selection step. In other words, they reduce the number of features that are fed into the classifier, thus reducing the computational complexity. The authors claim that this step also increases the classification accuracy. For this feature selection step, they iteratively select a feature set that maximizes a separability criterion. This criterion is a kernel-based class separability method, that was developed by Wang [197]. This method consists in projecting the samples to a kernel space and calculating the separability of the different features in that kernel space. Once the features have been selected, the final

step is to use a classifier to predict the emotion. In this case, the classifier employed is a least-square SVM.

3.2.1.2 Emotion Recognition From EEG Signals

Subramanian et al. [184] use different physiological signals to recognize high and low categories of arousal and valence. We review how they process EEG signals, as they are of our interest, but for the other signals they employ a similar procedure, although differing in the features that are extracted at the beginning of the process.

Subramanian et al. [184] employ 1-channel EEG signal, with this channel monitoring the frontal lobe activity. From this signal, they extract 88 features. Some examples of the extracted features are: mean of the signal, standard deviation, skewness, mean number of peaks, and others. Then, they use Fisher's linear discriminant [69] to identify the most discriminative features. Using the selected features, they train a Naive Bayes and a SVM classifier to perform recognition of high and low categories of arousal and valence.

3.2.1.3 Discussion of Classical Machine Learning Approaches

The above examples illustrate the general procedure for performing emotion recognition when using classical ML techniques: selecting and extracting different features from the signal, using an algorithm to reduce the number of features, and feeding those features into a ML classifier. This procedure requires that an expert decide which features will be employed. Even if the approach employs an automatic feature selection mechanism, the original features to be used as input for the system still need to be designed and selected by an expert.

Having an expert to select features has the advantage that in a way, external knowledge is included in the system, which can be beneficial for the task. On the other hand, it could be difficult to select the features that will lead to good performance. In fact, the work of Shukla et al. [173] shows that features that are commonly used to predict arousal and valence are not necessarily the most discriminant. For this reason, it is tempting to explore options where the features are extracted in a data-driven manner, so no domain knowledge is required. DL techniques can be used for this purpose, since they are capable of deriving features directly from data, with the model learning intermediate features that better suit the target task, thus leading to more accurate results [192]. Moreover, processing raw signals with DL

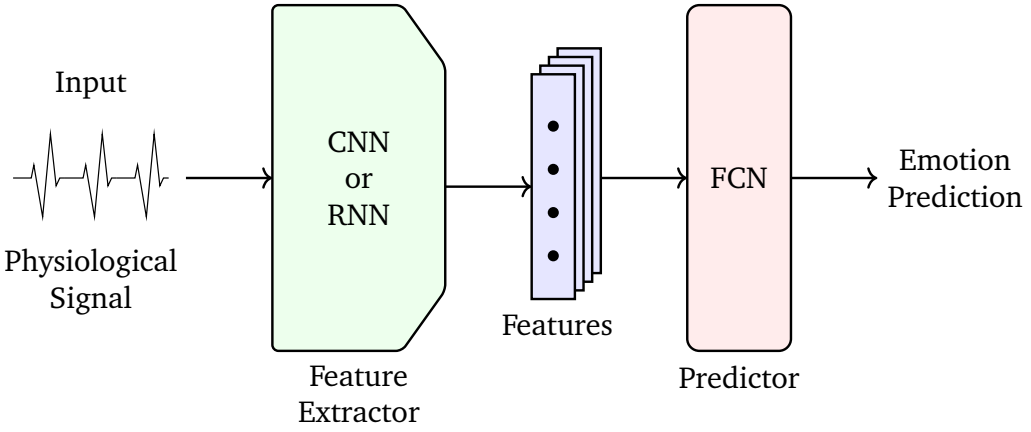


Figure 3.2 – DL approach for emotion recognition from physiological signals.

models allows them to benefit from certain classes of data transformations such as Fourier Transforms, Cepstral Coefficients or Gaussian Scale Space, as a pre-processing step to the model.

3.2.2 Deep Learning Approaches for Emotion Recognition

DL architectures have demonstrated that they are capable of automatically extracting useful features to perform a variety of tasks in computer vision, Natural Language Processing (NLP), signal processing, and other domains [47]. As opposed to engineered features, where expert knowledge is needed to choose which parameters to use, deep neural networks are capable of extracting the features directly from data. When using a DL approach, the extracted features may not be explicitly related to what is known about the problem, so in a sense, the model extracts useful information from the signal that may not be evident. Note that even though DL models can process raw data, it is still feasible to pre-process the input signals using operations such as a Fourier Transform to transform the input into a canonical representation, thus reducing the dimensionality of the input signal prior to processing int with the DL model.

Regarding emotion recognition from physiological signals, a common approach, illustrated in Figure 3.2, is to use Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) layers to process the signals and model their spatio-temporal information. After these layers, a Fully-Connected Network (FCN) is commonly used as classifier to predict emotions. In this case, the CNN and RNN layers are considered feature extractors, and the FCN is defined as the predictor model.

An example of a contribution that uses the approach described in the previous paragraph is the work of Santamaria-Granados et al. [161], where ECG and EDA signals are processed with a 1D Convolutional Neural Network (1D-CNN) obtaining features from those signals. Another example is the work of Alhagry et al. [6], which uses a Long Short-Term Memory (LSTM) network (a RNN designed to address the vanishing gradient problem [88]) to extract features from raw EEG signals. Another example is the work of Harper and Southern [85], where features are extracted by combining a 1D-CNN network concurrently with a LSTM network. An additional example is the work of Hu et al. [91], which employs CNN-based layers to extract spectrogram-like features from raw EEG signals, and then these features are processed with additional CNN layers to obtain the final features. In all these four examples, the extracted features are processed with a FCN used to predict the emotion.

To better illustrate DL approaches for emotion recognition, Sections 3.2.2.1 and 3.2.2.2 detail the contributions of Santamaria-Granados et al. [161] and Hu et al. [91], respectively, both contributions using DL models for emotion recognition.

3.2.2.1 Using DL Models to Recognize Emotions from ECG and EDA Signals

In their work, Santamaria-Granados et al. [161] use 1D-CNNs to extract features from ECG and EDA signals, and then process those features with a FCN to predict low and high categories of arousal and valence.

Santamaria-Granados et al. [161] use two approaches: in the first one, they use raw signals as inputs. For the second approach, they extract some parameters from the signals and use those parameters as inputs. Specifically, they detect the QRS peaks from the ECG signal and use them as inputs, and in a similar way, for the EDA signal they detect peaks and use them as inputs. In both approaches, they process the inputs with a 1D-CNN followed by a FCN. In their experiments, they obtain similar results with both approaches when predicting arousal and valence from ECG signals. When using EDA signals, the approach that uses the peaks as inputs obtains more accurate results. The authors also compare their results with results obtained using classical ML models like Naive Bayes, k-Nearest Neighbors, and Random Forest, which use engineered features as inputs. This comparison shows that their DL-based method obtains more accurate results.

Santamaria-Granados et al. [161] show that it is possible to use DL

models to recognize high and low categories of arousal and valence, employing as inputs physiological signals, either using the raw signals or using some parameters extracted from the signals. Moreover, their DL approach shows better accuracy than other approaches consisting of using engineered features and a non-deep ML model.

3.2.2.2 Using DL Models to Extract Features from EEG Signals for Emotion Recognition

Hu et al. [91] explore the use of a convolutional layer designed to extract in a data-driven manner spectrogram-like features from raw EEG signals, so those features can be used to recognize high and low categories of arousal, valence and dominance.

Their convolutional layer works by average-pooling the kernel used for convolution. Specifically, this pooling is done l times, scaling the kernel each time to a specific period, thus capturing specific frequency-like characteristics from the input signal. Each pooled kernel is used to do convolution with the input signal, thus at the end a 2D spectrogram-like feature of size $l \times \textit{signal length}$ is obtained. Finally, this 2D feature is processed with a 2D Convolutional Neural Network (2D-CNN) and then with a FCN to perform the classification.

The work of Hu et al. [91] is an example of how DL models can be used to extract features from raw physiological signals. Moreover, their results show an improvement over other contributions that use engineered features.

3.2.2.3 Discussion on DL Approaches

DL approaches have the advantage that can extract data-driven features from raw data, avoiding having to design and select engineered features, which may be difficult to do. Therefore, we are interested in a DL solution for emotion recognition, that uses raw physiological signals as input. On the other hand, in order to have a DL model to perform effectively, enough data is needed. We thus argue that the performance of emotion recognition models, for which there are no large labeled datasets, should improve if the problem of lack of data is addressed. Addressing this issue is in fact Challenge 3.1, which was identified in Section 3.1.2. We discuss how other contributions address this challenge in Sections 3.2.4 and 3.2.5.

Another point to note is that the majority of the reviewed approaches

employ CNN-based feature extractors. One drawback of these CNN-based approaches is that they do not take context into account: after training, kernel weights of the CNN are static, no matter the input. It is possible that the result could be improved by dynamically scoring the relevance of different parts of the input, as done by the attention layers of the Transformer [196], which is described in Section 2.3. This is in fact a way to address Challenge 3.2, i.e. how to use a DL model to extract suitable features. The following section provides a review of some works that use Transformer-based approaches to address this challenge for the task of emotion recognition.

3.2.3 Transformer-Based Emotion Recognition

DL architectures based on attention, such as the Transformer [196], can dynamically weigh the importance of different parts of the input. Although developed for NLP tasks, the Transformer has been successfully used in other domains like computer vision [52] and audio processing [14], demonstrating its versatility for different tasks.

Transformers have been used to process time-series for the task of time-series forecasting, like in the works of Li et al. [121] and Wu et al. [206], showing that Transformers are also useful for processing time series data. This is relevant because physiological signals can be seen as a type of time series. In fact, some authors have performed analysis of medical physiological signals using Transformers. For example, Ahmedt-Aristizabal et al. [4] use Transformers to analyze physiological recordings to recognize neurogenerative disorders, neurological status, and seizure type. Another example is the work of Yan et al. [211], where a Transformer is used to process ECG signals for heartbeat classification to help with the diagnosis of cardiac arrhythmia.

Regarding emotion recognition, several works use Transformers to perform this task. Many of these works deal with multimodal signals, combining text, visual, and audio information [95, 190, 208]. Other works also use physiological signals in addition to the text, visual and audio inputs [30, 41].

Using Transformers only with physiological signals for emotion recognition has been explored by some authors. For example, Arjun et al. [9] employ a type of Transformer, called the Vision Transformer [52], to process EEG signals. Another example is the work of Behinaein et al. [22], which uses ECG signals to detect stress. In this work, the signals are processed by a 1D-CNN, followed by a Transformer encoder, and then using a FCN as predictor.

For illustration purposes, the following subsection describes in detail the work of Arjun et al. [9].

3.2.3.1 Using Transformers to Recognize Emotions from EEG Signals

In [9], Arjun et al. design a Transformer-based model to recognize high and low categories of arousal and valence from EEG signals. To do this, they employ a variation of the Transformer, called Vision Transformer [52], to process the EEG signals. The original Vision Transformer [52] is designed for computer vision tasks, where its input sequence is formed by flattened patches from the input image, ordered sequentially.

Arjun et al. [9] present two approaches: one where the EEG signals are first converted to images using continuous wavelet transformations [5], and then feeding these images into the Vision Transformer; and another approach where raw EEG signals are used directly. For the second approach, the patches are segments from the raw signal.

The results from Arjun et al.'s work [9] show that using raw EEG signals obtains better performance in terms of accuracy than using EEG images obtained with continuous wavelet transformations. The work of Arjun et al. [9] demonstrates that it is feasible to use Transformers to process raw physiological signals, and that they are capable of extracting suitable features for emotion recognition.

3.2.3.2 Discussion on Transformer-Based Emotion Recognition from Raw Physiological Signals

Arjun et al.'s work [9] and other Transformer-based approaches named in this section are examples of the direction that we also take: using Transformers as the backbone to process raw physiological signals. Specifically, we find that the characteristics of the Transformer, such as having the capacity to account for long-range dependencies, should be useful when processing raw physiological signals for emotion recognition. Also, the attention mechanisms of the Transformer are capable of recognizing the important parts of the raw signals, giving more weight to those important parts when extracting the features.

However, the approaches presented in this section rely on supervised learning, and therefore, as mentioned before, they are limited by the availability of labeled training data. Several techniques have come to solve this issue, which are reviewed below.

3.2.4 Training DL Models with Limited Labeled Data

When a supervised DL approach is employed to process raw physiological signals, it is relaying on the capabilities of the DL model to find patterns present in the data to extract useful features. This typically requires having enough labeled data to exploit the full potential of a DL approach. The problem is that large datasets of physiological signals with labels of emotion are difficult to obtain, thus the size of the datasets might not be enough to exploit effectively a DL model.

A solution to obtain more accurate results from a model under data-constrained situations can be incorporating external knowledge during training. This can be done by employing knowledge and expertise about a task, and with this expertise designing and selecting different parameters from the data. In other words, engineered features can be employed. This approach is used by Santamaria-Granados et al. [161], where they use a CNN to process parameters extracted from physiological signals. This work was reviewed in Section 3.2.2.1.

An extreme case of limited labeled data is few-shot learning, where the aim is to train a model using few labeled examples per class, typically less than 5. Wang et al. [199] describe several techniques used in the literature to address the tasks of few-shot learning including data augmentation, where the available data is augmented by doing some transformations on it, thus creating more examples; embedding learning, where each sample is embedded to a lower dimensional field, such that the embeddings corresponding to the same class are closer while the embeddings corresponding to different classes are more separated, thus making easier to differentiate classes; and pre-training methods, where a model is pre-trained for another task, and then the parameters of the pre-trained model are fine-tuned for the target task.

In our case, we assume that the data available for training have more than a *few* examples for each class, which is a reasonable assumption if the different datasets of physiological data with labels of emotion are analyzed (see Appendix A). Therefore, we do not have a few-shot learning problem. Nevertheless, the techniques used to address the problem of few-shot learning can be adapted to obtain more accurate results from a DL model in data-constrained situations. Notably, pre-training techniques are widely used in different domains like computer vision, NLP, and signal processing, to improve the performance of models, especially when large amounts of unlabeled data (or large amounts of data labeled for a different

task) are available.

In this chapter, we are interested in pre-training a DL model, specifically a Transformer-based model, for the task of emotion recognition from raw physiological signals. We do this because we want to take advantage of the availability of large amounts of physiological data used for medical tasks. The data is especially abundant for ECG and EEG signals. Thus, our idea is to use these data, which does not necessarily contain labels of emotion, to pre-train a Transformer-based model, and then fine-tune this model to perform emotion recognition. Since we use a pre-training technique, in the following section we review this type of technique in detail, also reviewing other contributions that use this method.

3.2.5 Learning Pre-trained Models with Self Supervised Learning

Pre-training techniques are used to improve the performance of deep neural networks, and if done in a self-supervised fashion, it is possible to take advantage of unlabeled data that is typically more abundant than labeled data. Below, we provide a revision of the self-supervised pre-training technique, followed by a review of related works that use pre-training approaches for emotion recognition.

3.2.5.1 Self-Supervised Pre-Training

Pre-training is a technique employed to boost the performance of deep neural networks [185]. Moreover, as pre-training acts as a regularizer for the subsequent training [60], it is especially useful when training data is scarce since in this situation the model is prone to overfitting.

The process of pre-training and then fine-tuning a model consists of the following. First, the model is pre-trained using a pretext task, such that the model learns to generate good features for the actual final task. Then, the pre-trained model, now capable of generating good features, can be further trained for the target task, in a step called fine-tuning. These learned features are typically called *representations*, and thus we can refer to the process of pre-training a model to obtain these representations as *representation learning* [23]. In addition, when going from the pre-training phase to the fine-tuning phase, there is a *transfer* of what the model learned for the pretext task to the target task. For this reason, sometimes this part is called *transfer learning*.

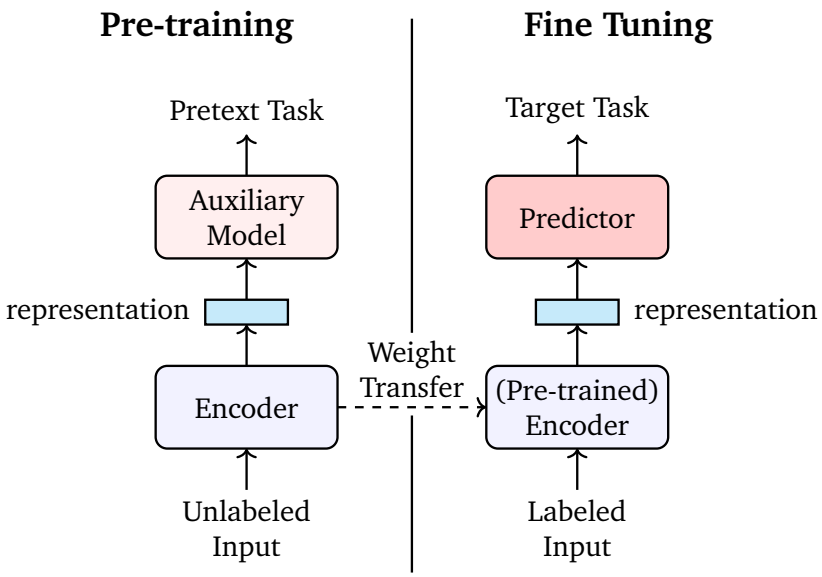


Figure 3.3 – Pre-training and fine-tuning a model

The pretext task should allow the model to learn representations that capture useful information from the signal. Ideally, this task should be *self-supervised*, meaning that no manual labels are required to perform the pretext task. This has the advantage of allowing the usage of unlabeled data during the pre-training phase, which is more abundant and easier to gather than labeled data. So, instead of using labeled data during pre-training, the idea is that the supervision comes from the unlabeled input data itself, by carefully designing the pretext task [61]; hence the term self-supervised. Accordingly, when pre-training a model using the self-supervised paradigm, we refer to this as *self-supervised pre-training*.

One example of a self-supervised pretext task is reconstructing the original input. For instance, if processing images, during pre-training the model will output a representation of this image. Then, the image is reconstructed from this representation with the aid of the auxiliary model. The pretext task can be learned by comparing the reconstructed image with the original one. Therefore, no manual labels are needed for this, but note that the task can still be considered as supervised.

Figure 3.3 illustrates a possible approach for pre-training and fine-tuning a model. Three phases can be identified: pre-training, weight transfer, and fine-tuning. First, during the pre-training phase, left part of Figure 3.3, the model processes the unlabeled inputs to generate the representations. The module that generates these representations is called *encoder*, since

it encodes the inputs into representations. These representations are then further processed with an auxiliary model to accomplish the pretext task.

Next, the weight transfer is performed. When the pre-training is finished, the auxiliary model is discarded, and the pre-trained encoder will be used to solve the target task. Another way of seeing this is having a new encoder with exactly the same architecture as the original encoder and transferring the weights from the pre-trained encoder to the new one, as indicated by the arrow marked as "Weight Transfer" in Figure 3.3.

Finally, the model is fine-tuned, where the model is trained to perform the target task, using the labeled inputs. For this, a new network is added to process the representations generated by the encoder, with the objective of accomplishing the target task, as depicted by the right part of Figure 3.3. In this figure, this new network is called predictor, since it typically predicts a class or a value. During the fine-tuning step, the pre-trained encoder might be frozen and only the predictor network is trained; or the complete architecture, the predictor network and the pre-trained encoder, might be trained.

Three types of self-supervised pre-training approaches can be distinguished, according to the pretext task. Summarizing the work of Del Pup and Atzori [49], those types are:

1. **Predictive pretext tasks**: The pretext task consists of regression or classification problems. For example, the input can be transformed by adding noise, scaling it, flipping it, etc. Then, the pretext tasks consist of classifying which transformation has taken place.
2. **Generative pretext tasks:** The pretext task consists in regenerating the input data from a corrupted version of the data. For example, if processing sentences, we can mask some words in the input sentences, and the pretext task may consist in predicting those missing words.
3. **Contrastive Learning pretext tasks:** The idea behind this type of task is that representations should be closer if they come from related inputs, and farther if they come from unrelated inputs. One way of doing this is by generating new inputs by transforming the original input by adding noise, scaling it, flipping it, etc. Then, related inputs are transformed inputs that come from the same sample, while unrelated inputs come from different samples. To measure how close or far the representations are, cosine similarity can be used. In summary, in this case the pretext task consists in measuring the distance between a pair of representations, pushing them closer if they come from the

same transformed sample, and pushing them away if they come from different samples.

3.2.5.2 Pre-Trained Models for Emotion Recognition

In the literature, we found that authors have used the different types of pretext tasks identified in Del Pup and Atzori [49], and mentioned above, namely predictive tasks, generative tasks, and contrastive tasks. Below, we review different works that use these approaches for the task of emotion recognition.

Using a Contrastive Pretext Task

Kan et al. [98] and Shen et al. [169] use a contrastive learning approach to pre-train a model for emotion recognition from EEG signals. In these works, the authors take advantage of the fact that in order to obtain the emotion-induced EEG signals, researchers who build datasets usually employ the same stimuli, typically a video, in several subjects. Consequently, in the datasets usually exist samples obtained from different subjects, but with these subjects having received the same stimuli. Then, the general idea behind the contrastive task in the works of Kan et al. [98] and Shen et al. [169] is to maximize the similarity of the representations from EEG samples that were triggered by the same stimuli.

Using a Generative Pretext Task

Ross et al. [154] pre-train a model using a generative approach. In their work, they use an autoencoder [18], encoding and then reconstructing the signal during the pre-training phase. Specifically, the autoencoder generates a representation from the input signal, and then the signal is reconstructed from this representation. The model can be pre-trained by comparing the original signal with the generated one. Since the dimension of the representation is smaller than the dimension of the input signal, this representation should contain the most important information in order to reconstruct the input successfully. Therefore, this approach leads to the generation of strong representations.

Using a Predictive Pretext Task

Sarkar and Etemad [162], address the task of emotion recognition from raw ECG signals. To pre-train their model, the authors use a predictive approach. The pretext task consists in applying transformations to the input signal, and then recognizing what transformation was used. Specifically, they use 6 different transformations:

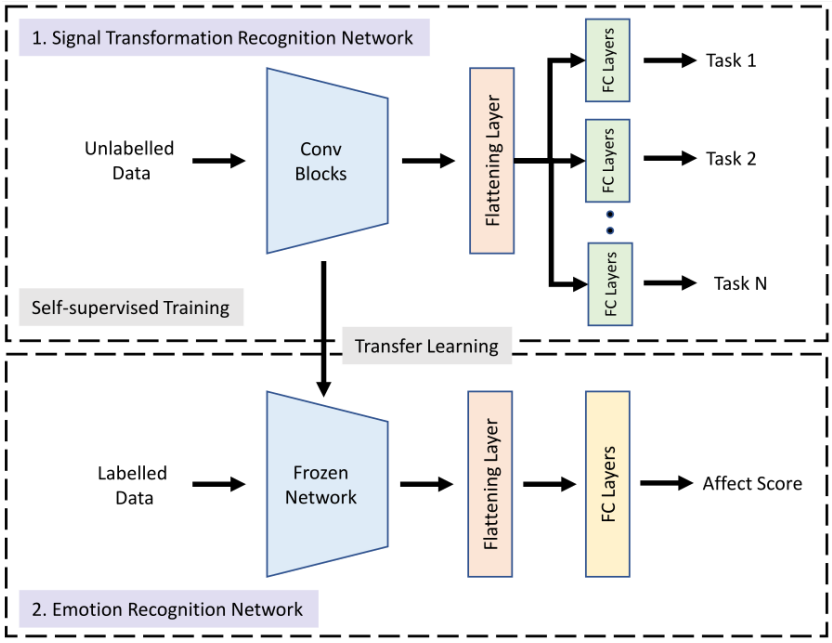


Figure 3.4 – Depiction of the approach used by Sarkar and Eteman [162]. Affect Score refers to the predicted value of arousal or valence. (Figure from their related paper [163]).

— *Noise Addition:* Random noise from a Gaussian distribution is added to the signal.

— *Scaling:* Each signal value is multiplied by a scaling factor.

— *Negation:* The signal values are multiplied by -1.

— *Temporal Inversion:* If we represent the signal as $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$, the signal is transformed to the sequence $\{x_n, x_{n-1}, \ldots, x_2, x_1\}$.

— *Permutation*. The original signal is divided into segments, and the transformed signal is the result of shuffling those segments.

— *Time-warping*: Randomly selected segments of the input signal are squeezed or stretched along the time axis.

Figure 3.4 depicts the model used by Sarkar and Etemad. This figure shows that the signal is first processed by CNN layers. Those are the layers that are also used for the target task after the pre-training phase. Thus, the authors call these layers *shared layers*. For the pretext task, an auxiliary network consisting of 7 branches of FCN is added to the shared layers. The first 6 branches are used to predict if each transformation has taken place, and the last one predicts if no transformation was used. After the

pre-training phase, the auxiliary network is discarded, and a new FCN is added to the shared layers to fine-tune the model for emotion recognition.

3.2.5.3 Discussion on Pre-Trained Models

In the preceding subsection, we have reviewed several works that use pre-training for the task of emotion recognition from physiological signals. All the contributions that were presented are based on CNNs. As we have already mentioned, we believe that a Transformer-based approach can be advantageous. In addition, although some works have used pre-training techniques with Transformers to process time-series [84, 215], none of these works deal with uni-modal physiological signals for emotion recognition.

3.2.6 Discussion

This section reviewed several contributions that address the task of emotion recognition. Approaches that use engineered features have the advantage that external knowledge is incorporated into the approach. On the other hand, selecting relevant features can be a difficult task. DL approaches are useful for this situation since they extract features in a data-driven fashion. Moreover, DL approaches have the advantage that they can be pre-trained, as has been done by several authors [98, 154, 162], to further improve the accuracy of the results obtained from a model.

Transformers have been shown to be useful for tasks of emotion recognition from physiological signals. This architecture presents several advantages, like its capacity to model long-range dependencies, and to attend (i.e. give more weight) to the important parts of the input signal. Moreover, we believe that employing self-supervised techniques to pre-train a Transformer should help the architecture to produce stronger features from physiological signals, leading to more accurate results when predicting emotions. Consequently, we orient our contribution towards this idea. To the best of our knowledge, we are exploring a novel approach by examining the potential of pre-training a Transformer model for recognizing emotions from ECG and EEG signals.

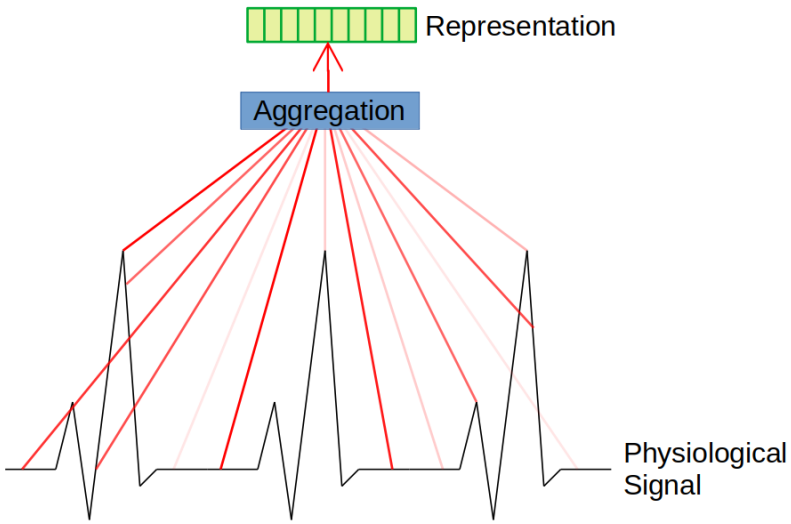


Figure 3.5 – Generating the representation of a signal: the representation is built taking into account the importance of different parts of the signal.

3.3 Pre-Trained Transformer for Emotion Recognition

This section describes our contribution to the problem defined in Section 3.1: recognizing high and low categories of arousal and valence from raw physiological signals. Our solution needs to address the Challenges 3.1 and 3.2 described in Section 3.1.2, namely processing raw physiological signals effectively, and not having large quantities of labeled data. For simplicity, for the rest of the chapter, the term emotion recognition refers to recognizing high and low categories of arousal or valence.

To address the challenge of processing the raw signals effectively, we design a model with the Transformer [196] as the backbone of our solution, justifying this decision in Section 3.3.1, and presenting the details of our Transformer-based architecture in Section 3.3.2. To address the challenge of not having large quantities of labeled data, we develop a self-supervised pre-training technique, which is explained in Section 3.3.3, followed by an explanation of the fine-tuning step in Section 3.3.4.

3.3.1 Transformers for Raw Physiological Signals

Since we want to process raw signals, it is useful to aggregate information from the whole signal, giving more weight to more important parts of the signal. A way of doing this is using attention mechanisms that weigh, or pay more attention, to the relevant parts of the input. The Transformer, described in Section 2.3, is currently the most successful attention-based approach, and thus we base our approach on this architecture.

A depiction of using attention to build a representation of the signal is shown in Figure 3.5. The contribution of each part of the signal to the representation is determined by an attention score, illustrated with the opacity of the red lines. The Transformer should be capable of assessing this importance when building the representation.

While the Transformer was originally developed for NLP tasks involving transformation of sequences of symbols (words and phrases), it is based on encoders and decoders that transform vectors of numerical values. Applying a Transformer to natural language requires transforming the symbolic input into sequences of vectors of numerical values using an embedding. Physiological signals are sequences of numerical values that can also be transformed into sequences of numeric vectors that can be processed by a Transformer.

3.3.2 Our Architecture

Our Transformer-based approach used to obtain representations from raw physiological signals is depicted in Figure 3.6. Our approach consists of two phases. First, we pre-train the model, illustrated by the left part of Figure 3.6. Second, we fine-tune the model, depicted in the right part of Figure 3.6.

As mentioned in Section 2.2.5, we use the same architecture to recognize arousal and valence, and train different models for each of those emotion dimensions. Specifically, a single model is pre-trained for both arousal and valence recognition, but one model is independently fine-tuned for arousal and another model for valence. This approach is a good trade-off between designing two specialized systems, and having a general model that might sacrifice individual performance. In fact, several other works also use a single architecture trained independently to recognize the different emotion dimensions [91, 161, 162].

Our architecture generates representations with a component that we

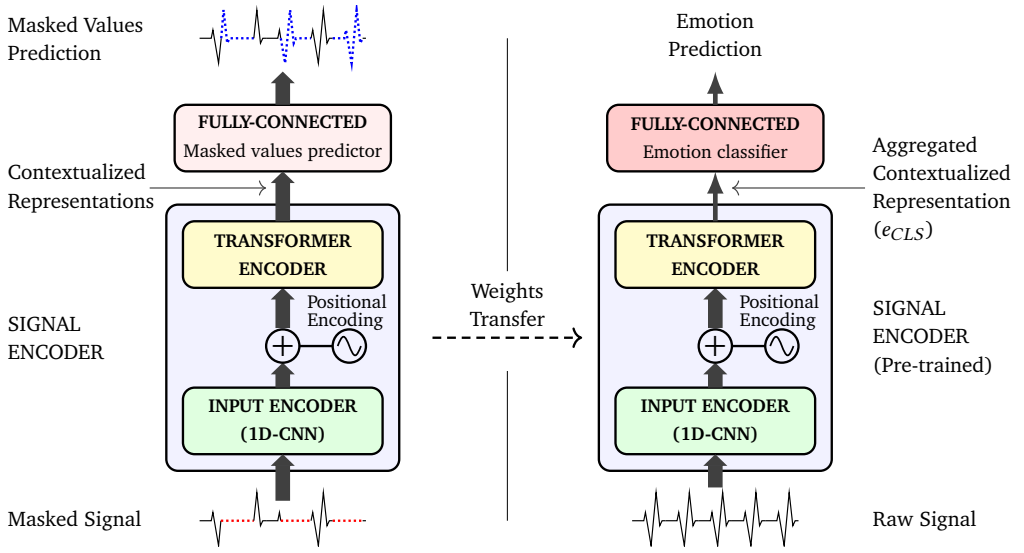


Figure 3.6 – Our approach to pre-train (left) and fine-tune (right) a Transformer to process raw physiological signals.

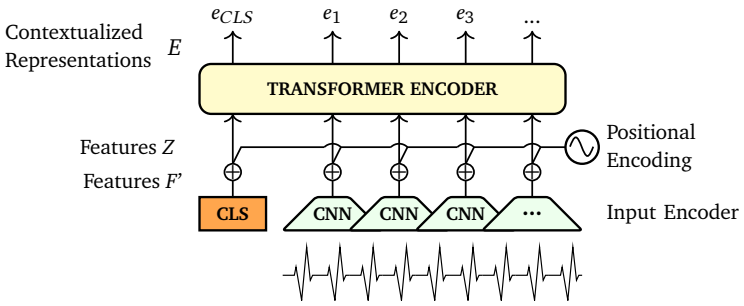


Figure 3.7 – Our Transformer-based signal encoder to generate representations. The aggregated representation e_{CLS} is used for classification.

refer to as *signal encoder* in Figure 3.6. We now explain each of the components of the signal encoder, which is depicted in Figure 3.7.

3.3.2.1 Input Encoder

In order to process the raw physiological signals with our signal encoder, it is necessary to first encode these signals into s feature vectors of dimension d_{model}, getting one vector for each of the s values of the input signal. A 1D-CNNs can be used as input encoder, this way the input encoder will aggregate local information [38].

We represent each scalar value of the raw input signal as x_i, then the

input signal of length s is

$$X = \{x_1, ..., x_s\}. \tag{3.1}$$

We then encode the signal with the 1D-CNN to obtain the features $f_i \in \mathbb{R}^{d_{\text{model}}}$, so at the end of the input encoder we have

$$F = \{f_1, ..., f_s\} = \text{1D-CNN}\left(\{x_1, ..., x_s\}\right). \tag{3.2}$$

3.3.2.2 CLS Token

The target outputs are low and high categories of arousal and valence, based on the entire input sequence, so it is necessary to obtain a single representation of the whole input signal. This is provided by appending a special token to the beginning of the feature sequence F, as done by Devlin et al. for the BERT model [51]. This token is called *classification token* or CLS for short. Thus, after adding the CLS token, the sequence of features becomes

$$F' = \{CLS, f_1, ..., f_s\} = \{CLS; F\} \tag{3.3}$$

where ; denotes concatenation.

In our case, the CLS token is a learnable vector with dimension $\mathbb{R}^{d_{\text{model}}}$, that is trained with the rest of the model. As we shall see, once the sequence is processed by the Transformer, we obtain the representation of the CLS token (e_{CLS} in Figure 3.7). The attention mechanisms of the Transformer allow that the information from the entire input signal to be aggregated in e_{CLS}. Recall that the Multi-Head Attention (MHA) mechanism from the Transformer, described in Section 2.3, has as inputs a query Q, a key K, and a value V. When the CLS token is processed, it becomes the Q that queries the keys from all the other values of the input signal, weighting these other values and effectively incorporating their information in the representation generated from this token. Therefore, when performing the classification, e_{CLS} can be used as input for the classifier network.

3.3.2.3 Positional Encoding

As discussed in Section 2.3, Transformers are permutation-invariant, and information about the order of the input values has to be injected explicitly. In our case, we use fixed sinusoidal positional embeddings as proposed by Vaswani et al. [196]. We add these positional embeddings to the features

F' and then apply layer normalization LN [11] to the resulting vector. If $pe_i \in \mathbb{R}^{d_{\text{model}}}$ is the positional embedding for the time-step i, we have

$$Z = LN(\{CLS + pe_0, f_1 + pe_1, ..., f_s + pe_s\}), \tag{3.4}$$

where Z is the sequence of features that will be provided to the encoder.

3.3.2.4 Transformer Encoder

Since this part of the architecture encodes the signals to obtain representations, we use the encoder part of the Transformer discarding the decoder part. Specifically, we use a Transformer encoder to obtain the representations E. The Transformer encoder is composed by a MHA module followed by a fully-connected Feed-Forward Network (FFN). Section 2.3 gives more details about the Transformer encoder. Thus, to process the features Z we have

$$E = \{e_{CLS}, e_1, ..., e_s\} = \text{Transformer\_Encoder}(Z). \tag{3.5}$$

The representations E, specifically e_{CLS}, are used to perform the emotion recognition, as is described in Section 3.3.4.

3.3.3 Pre-Training the Signal Encoder

We use an approach inspired by the BERT model presented in Devlin et al. [51], to pre-train the signal encoder in a self-supervised fashion. First, we mask random segments of a certain length in the input signal. This is done by replacing the masked values with zeros. Then, we train the model to predict those masked values. This process is depicted in the left part of Figure 3.6. Note that for this step, no labeled data is needed. Although we use zeros to mask segments of the signal, it is possible to use other values for masking, like a value outside the range of the signal, for example. We do not study the impact of the values used for masking, leaving this as a perspective to explore.

To mask the signal, we follow an approach similar to the one described in Baevski et al. [14]. First, a proportion p of points is randomly selected from the input signal. These selected values become the starting points of the masked segments. Then, for each starting point, the subsequent M points are masked. There might be overlapping between the masked segments. Figure 3.8 depicts our masking strategy.

A FCN is used to predict the masked points. The FCN is placed on top of the signal encoder, as depicted in the left part of Figure 3.6. We do not

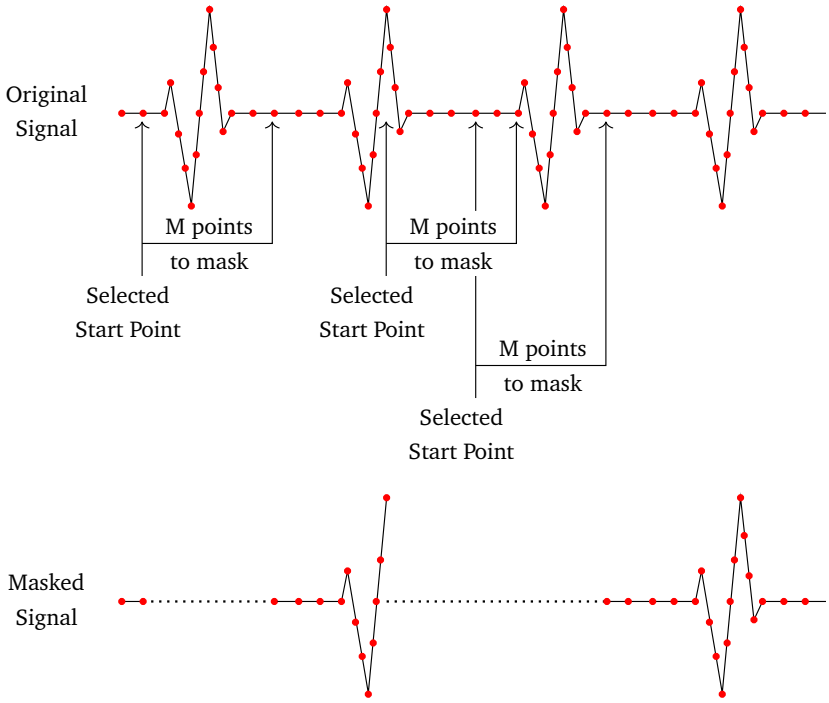


Figure 3.8 – Masking strategy. We randomly selected points as the starting points of segments of length M. The masked values are replaced with zeros.

reconstruct the complete signal. Instead, we predict only the values of the masked inputs.

Since the model predicts values for the masked parts of the signal, we want to minimize the difference between the predicted values and the real (original unmasked) values. Thus, for this pre-training phase, the reconstruction loss $\mathcal{L}_r$ is the mean square error between predicted and real values:

$$\mathcal{L}_r = \frac{1}{N_m} \sum_{j=1}^{N_m} (\hat{x}_j - x_{p(j)})^2, \tag{3.6}$$

where N_m is the number of masked values, $\hat{x}_j$ is the prediction corresponding to the j^{th} masked value, and $x_{p(j)}$ is the original input value selected to be the j^{th} masked value, whose position is $p(j)$ in the input signal.

3.3.4 Fine-Tuning the Model

To fine-tune the model to recognize high and low categories of arousal and valence, a supervised approach is used. This means that now labeled

data is employed. As shown in the right part of Figure 3.6, a FCN is added on top of the signal encoder, replacing the network used as masked-values predictor. This new FCN, which works as the emotion classifier, receives as input e_{CLS}.

The new FCN is randomly initialized, while the signal encoder is initialized with the weights obtained after the pre-training phase. During this phase, all the parameters of the model, including the pre-trained weights, are fine-tuned. We use the binary cross-entropy loss $\mathcal{L}_{ft}$ as the fine-tuning loss:

$$\mathcal{L}_{ft} = -\sum_{n=1}^{N} \left(w_p y_n \log(\sigma(out_n)) - (1 - y_n) \log(1 - \sigma(out_n)) \right), \tag{3.7}$$

where y_n is an indicator variable with a value of 1 if the class of the ground truth for sample n corresponds to a high level of arousal (high emotion intensity) or high level of valence (positive emotion), and 0 if it corresponds to a low level of arousal (low emotion intensity) or low level of valence (negative emotion). N is the number of samples in the minibatch, out_n is the output of the classifier for sample n, and σ is the sigmoid function. We use the ratio of negative to positive training samples w_p to compensate for the unbalances that may be present in the dataset.

3.3.5 Expected Results

We believe that using our pre-training strategy, the model should be able to generate stronger representations from the data, and be less prone to overfitting. Therefore, we expect the following result regarding recognizing emotions from physiological signals:

Expected Result 3.1. *Using our strategy to pre-train a model and then fine-tune it to recognize high and low categories of arousal and valence, should give more accurate results than training a model from scratch.*

In addition, we believe that a Transformer-based architecture should be more suitable to model the dependencies from the inputs than other DL networks like CNNs or RNNs, when processing physiological signals for emotion recognition. This, combined with our pre-training strategy, led us to expect the following result:

Expected Result 3.2. *Our pre-trained Transformer-based approach should improve the results in terms of the accuracy of predictions of arousal and valence, compared to other state-of-the-art methods.*

Nº	Evaluation Dataset	Physiological Signal
1	AMIGOS	ECG
2	DREAMER	ECG
3	AMIGOS	EEG
4	DREAMER	EEG

Table 3.1 – Evaluation scenarios to test our approach for emotion recognition

Finally, we use a Transformer-based approach because its attention layers can identify and give more weight to the important parts of the physiological signal used as input. Then, we expect the following:

Expected Result 3.3. *The attention layers of the Transformer will give more attention to certain parts of the input signal, thus demonstrating that using this architecture is valuable for physiological signals.*

3.4 Experiments

This section presents the evaluation of our Transformer-based approach designed to recognize high and low categories of arousal and valence from raw physiological signals. Our approach is tested with ECG and EEG signals. This section starts by describing our experimental setup in Section 3.4.1, followed by the presentation of the results obtained when our approach is evaluated in Section 3.4.2.

3.4.1 Experimental Setup

3.4.1.1 Evaluation Scenarios

To evaluate our approach, we need data to fine-tune and test the model. We call these data the *evaluation dataset*. We also need data to pre-train the model, which we call the *pre-training datasets*. We call an *evaluation scenario* assessing the model with an evaluation dataset using a specific physiological signal. Table 3.1 describes the evaluation scenarios for our experiments. The datasets used are the AMIGOS [135] and DREAMER [102] datasets, which provide ECG and EEG signals collected from subjects who watched videos specially selected to evoke an emotion. In both datasets, each subject conducted a self-assessment of their emotional state after watching each video, rating their levels of arousal and valence on a scale of 1 to 9 in

AMIGOS, and on a scale of 1 to 5 in DREAMER. In the AMIGOS dataset, there are data from 40 participants and around 65 hours of data were collected. For the DREAMER dataset, data comes from 23 subjects, and around 23 hours of data were collected. Appendix A provides more details about these datasets.

When processing EEG signals, the following 10 channels are used: F7, F3, T7, P7, O1, O2, P8, T8, F4, F8. We use these channels as they cover different regions around the entire scalp (see Figure 2.8b). Moreover, they are commonly present in the datasets that we use for pre-training.

Regarding ECG signals, the datasets provide signals taken from the left and right sides of the body. From preliminary experiments and observations, we did not find a difference between using any of those channels. For our experiments, we use the signals taken from the left side.

3.4.1.2 Datasets for Pre-Training

To pre-train the model using a self-supervised approach, we gather datasets that include the necessary physiological signals. It is important to note that these data do not need to contain labels of arousal and valence. We collect two sets of data: One with ECG signals and one with EEG signals.

ECG Data. The following datasets are used: ASCERTAIN [184], PsPM-FR [194], PsPM-HRM5 [145], PsPM-RRM1-2 [12], and PsPM-VIS [209]. We also employ the AMIGOS dataset and the DREAMER dataset as pre-training datasets. Since AMIGOS and DREAMER are also used as evaluation datasets, care is taken not to test and pre-train the model with the same samples. In order to obtain as much data as possible, we use all the ECG channels available in the datasets. The authors of the ASCERTAIN dataset provide a quality evaluation of the data. This evaluation is used to discard the signal that has a quality level of 3 or worse. The total amount of ECG data used for pre-training is around 230 hours.

EEG Data. The following datasets are used: WAY-EEG-GAL [129], BCI2000 [78, 164], and Large-EEG-BCI [103]. These datasets were collected to develop Brain-Computer Interfaces. We also use the AMIGOS dataset as part of the pre-training datasets. Since AMIGOS is also used as evaluation dataset, we pay attention to not use the same samples to pre-train and test our approach. Around 195 hours of EEG data are gathered to pre-train the model.

3.4.1.3 Signal Pre-Processing

Since our objective is to use raw physiological signals, we employ minimal pre-processing on those signals. The same pre-processing is done for the pre-training and evaluation data. We use an 8^{th} order Butterworth band-pass filter, with cut-off frequencies of 0.8Hz and 50Hz. This Butterworth filter was selected through preliminary experiments. The 50Hz cut-off frequency is used to eliminate the line noise, that is the noise that comes from the power source. The 0.8Hz cut-off frequency is used to eliminate any drift that the signal might have, i.e. slow variations of the signal that might be produced by movements of the sensor, for example. After filtering the signals, the next step is to down-sample them to a sample rate of 128Hz, if they have a sample rate greater than that. This way, all the signals will have a common sample rate. Next, we normalize the signals such that for each subject, they have zero-mean and unit-variance. Similar to other works that use physiological signals to predict emotions [135, 162, 154], we segment the signals and use each segment as a sample. We use 10-second segments as samples in our experiments as ten seconds is a short enough length to work efficiently with our Transformer-based solution, and we argue that it is long enough to capture an emotional response.

3.4.1.4 Pre-Training Set-Up

To maximize the quantity of data used to pre-train the model, we consider the AMIGOS and DREAMER datasets as part of the pre-training datasets. As these datasets are also used for evaluation, care was taken to avoid using the same samples to pre-train and test the model, while using as much data as possible from the gathered datasets to pre-train the model. To accomplish this, a different model is pre-trained for each evaluation scenario using different parts of AMIGOS and DREAMER datasets, and leaving for evaluation the samples not used in pre-training. A detailed explanation of how this is performed is provided in Appendix B.

To prepare the signals for our pretext pre-training task of predicting masked segments, for each of the 10-second segments, 3.25% of points are randomly selected as the starting points of masked spans of length $M = 20$. Since we allow overlapping of masked segments, on average this results in masking around 47% of each segment.

For all the evaluation scenarios, we build the input encoder of our model using a 1D-CNN composed of 3 layers, employing the Rectified Linear Unit (ReLU) activation function. At the first layer and at the output of the encoder,

we use layer normalization [11]. Each layer has a number of channels of 64, 128, and 256, with kernel sizes of 65, 33, and 17. The stride for all layers is 1. This configuration gives a receptive field of 113 input values, which at a sampling rate of 128Hz, is equivalent to 0.88 seconds. This receptive field size was selected because it matches the typical interval between peaks in an ECG signal, which has values between 0.6 seconds and 1 second, including when the person is experiencing emotions [207]. Based on preliminary experimental studies, we estimate that this receptive field is also suitable for EEG signals.

The Transformer in the signal encoder has a model dimension of $d_{\text{model}} = 256$. This Transformer has two layers and two attention heads, and the size of the FFN is $d_{\text{model}} \cdot 4 = 1024$. To predict the masked values, we use a single-layer FCN with ReLU activation function. The size of this FCN is $d_{\text{model}}/2 = 128$. When processing ECG signals, we employ an additional linear layer to project the output to a single value. This single value corresponds to the prediction of a masked point. Likewise, when processing EEG signals, we employ an additional linear layer to project the output to 10 output values. Each of these output values corresponds to the predicted value of each masked EEG channel.

For each evaluation scenario, the corresponding models are pre-trained for 500 epochs. We use learning rate warm-up, gradually increasing it during the first 30 epochs from $3.33e^{-5}$ to 0.001 when using ECG signals and to 0.0005 when using EEG signals. Then, the learning rate is linearly decreased. We use Adam's optimization with hyper-parameter values of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and L_2 weight decay of 0.005. A dropout of 0.1 is applied at the end of the input encoder, after the positional encoding, and inside the Transformer.

Bayesian Optimization with Hyperband (BOHB) [66] is used to tune the number of layers and heads of the Transformer, the learning rate, and the warm-up duration. This tuning is done using the Ray Tune framework [122].

3.4.1.5 Fine-Tuning Set-Up

During this phase, we train our model for emotion recognition in a supervised fashion. Specifically, we fine-tune our pre-trained model to predict high and low categories of arousal and valence, using the ECG and EEG signals in AMIGOS and DREAMER datasets. We fine-tune the complete model: the signal encoder and the FCN classifier.

Dataset	Signal	Learning Rate	Learning Rate Decay	Classifier Layer Sizes	Classifier Dropout
AMIGOS	ECG	0.0001	0.65 every 45 epochs	1024, 512	0.3
DREAMER	ECG	0.0001	None	128	0.2
AMIGOS	EEG	0.0001	0.65 every 45 epochs	64	0.6
DREAMER	EEG	0.0001	0.65 every 45 epochs	64	0.6

Table 3.2 – Hyperparameters used to fine-tune the models under the different evaluation scenarios.

For labels, we employ the emotional self-assessments provided in both datasets. Since we are interested in predicting high and low categories of arousal and valence, it is necessary to process the numerical values given in the self-assessments. In the AMIGOS dataset, the self-assessments provide numerical values of arousal and valence in the range of 1 to 9. In the DREAMER dataset, the range for arousal and valence is 1 to 5. To obtain the high or low labels from the numerical values, we find the average arousal and valence value in each dataset and use it as a threshold value to determine a low or high category of arousal and valence.

The FCN used as classifier uses ReLU as activation function. We add an additional linear layer at the output of the classifier to project the output to a single value. For each evaluation scenario, we fine-tune one model to predict arousal and another to predict valence. The models are fine-tuned for 100 epochs, using Adam optimization, with $\beta_1 = 0.9$, $\beta_2 = 0.999$. As it was done for pre-training, we use a dropout of 0.1 at the end of the input encoder, after the positional encoding, and inside the Transformer. Table 3.2 summarizes the rest of the hyper-parameters. As with pre-training, we use the Ray Tune Framework with BOHB to tune the different parameters of our model.

To evaluate our approach, we use 10-fold cross-validation, taking care of not using the same samples that were used for pre-training to evaluate the model. Details about this are given in Appendix B.

	Arousal Acc.	Arousal F1	Valence Acc.	Valence F1
Aggregation Method				
Last Representation	0.85±1.3e^{-2}	0.84±1.2e^{-2}	0.80±7.6e^{-3}	0.80±8.0e^{-3}
Max-Pooling 1	0.85±6.6e^{-3}	0.84±6.4e^{-3}	0.78±6.5e^{-3}	0.78±6.6e^{-3}
Max-Pooling 2	0.86±7.4e^{-3}	0.84±7.3e^{-3}	0.80±6.3e^{-3}	0.80±5.9e^{-3}
Average-Pooling 1	0.87±8.3e^{-3}	0.87±7.3e^{-3}	0.82±6.2e^{-3}	0.82±6.7e^{-3}
Average-Pooling 2	**0.88**±4.4e^{-3}	**0.87**±4.6e^{-3}	**0.83**±6.4e^{-3}	**0.83**±6.6e^{-3}
CLS	<u>0.88</u>±5.4e^{-3}	<u>0.87</u>±5.4e^{-3}	<u>0.83</u>±7.8e^{-3}	<u>0.83</u>±7.4e^{-3}
Segment Length				
40 seconds	0.86±1.2e^{-2}	0.85±1.1e^{-2}	0.82±1.0e^{-2}	0.81±9.9e^{-3}
20 seconds	<u>0.87</u>±5.6e^{-3}	<u>0.86</u>±6.4e^{-3}	<u>0.82</u>±7.8e^{-3}	<u>0.82</u>±8.1e^{-3}
10 seconds	**0.88**±5.4e^{-3}	**0.87**±5.4e^{-3}	**0.83**±7.8e^{-3}	**0.83**±7.4e^{-3}

Table 3.3 – Comparison of different strategies of our approach on the AMIGOS dataset with ECG signals for arousal. Best results are in bold, second bests are underlined.

3.4.2 Results

3.4.2.1 Metrics

To evaluate our results, we use as metrics the mean accuracy and the mean F1-score between the two predicted classes (high and low categories of arousal or valence), averaged across the 10 folds of cross-validation. We also report the confidence intervals of each metric, computed across the 10 folds of cross-validation. These confidence intervals are calculated using a t-distribution with 9 degrees of freedom for a two-sided 95% confidence. Specifically, the following expression is used to calculate the Confidence interval CI for each metric:

$$\text{CI} = \pm 2.262 \frac{S}{\sqrt{10}}, \tag{3.8}$$

where S is the standard deviation of the 10 results corresponding to each fold.

3.4.2.2 Preliminary Studies for the Aggregation Method and Segment Length

To test our model, it is necessary to determine an aggregation method and the segment length. The aggregation method concerns how to obtain a single representation from all the outputs of the signal encoder in order to make the prediction. As indicated in Section 3.3.2, we use the representation e_{CLS} of the *CLS* token as the aggregated information of the processed

segment. Nevertheless, other options can be used for that purpose; thus, we experimentally test those options and compare the results with using e_{CLS}. Second, regarding the segment length, we compare different segment lengths used to divide the input signal, in order to experimentally justify our choice of 10-second segments. We test these different options only in the evaluation scenario with AMIGOS as the evaluation dataset and using ECG signals. We expect the results to generalize to the other scenarios. The results of these experiments are reported in Table 3.3 and discussed below.

Aggregation Method: We compared several strategies to aggregate the representations given by the signal encoder. Note that the goal is to obtain a single vector to feed our FCN classifier. Below we describe the different strategies tested to get this aggregated vector.

— **CLS:** This is the strategy described in Section 3.3.2, where we use the representation of the *CLS* token, i.e. we use e_{CLS}.

— **Last Representation:** We use the last representation given by the signal encoder, i.e. we use e_s (see Expression 3.5).

— **Max-Pooling 1:** We apply max-pooling across all the output representations given by the signal encoder.

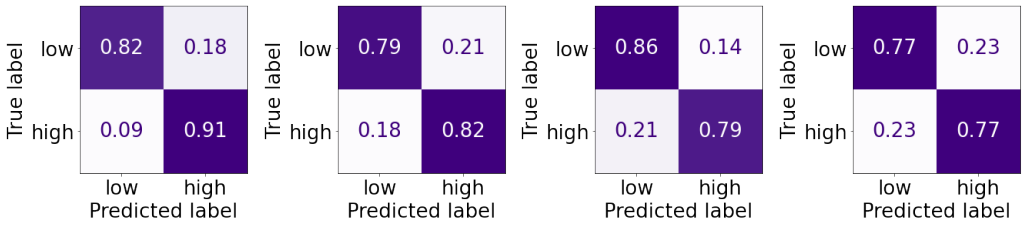
— **Max-Pooling 2:** We optimize a max-pooling strategy on the validation set: we reduce the representations to a size of 64, divide them into two groups, and then we apply max-pooling on each group. Finally, the results are concatenated to obtain a single representation of size 128.

— **Average-Pooling 1:** We apply average-pooling across all the output representations given by the signal encoder.

— **Average-Pooling 2:** We optimize an average-pooling strategy on the validation set: we divide the representations into 4 groups, and then we apply average pooling on each group. The next step is to concatenate the results to obtain a single representation of size 1024.

Table 3.3 shows that the best results are obtained using the Average-Pooling 2 strategy and using CLS. Even though the results are practically identical for both of them, the CLS strategy has the advantage of not requiring any kind of tuning on the validation data, as opposed to Average-Pooling 2. In fact, using the CLS token is a commonly-used strategy for Transformers. Therefore, for the remainder of the experiments, we will use CLS as our aggregation method.

Segment Length: We test 3 different segment lengths for dividing the physiological signals into the input samples. Specifically, we test segments



(a) Arousal (ECG) (b) Arousal (EEG) (c) Valence (ECG) (d) Valence (EEG)

Figure 3.9 – Confusion matrices with normalized rows, obtained with our approach using the CLS token, and 10-second segments as inputs. Results on the AMIGOS dataset, aggregating the results of the 10 folds.

of 10, 20, and 40 seconds. Table 3.3 shows that, for both valence and arousal, shorter segments lead to better results. We believe this is the case because longer segments should require more complex models, that is, a bigger Transformer and FCN classifier. These bigger models are harder to train due to the relatively low amount of labeled data on our evaluation datasets. Another advantage of shorter segments is that they are faster to process, permitting a higher number of training epochs and smaller learning rates. Thus, for the following experiments, we use 10-second segments.

Figure 3.9 shows the confusion matrices obtained in the AMIGOS dataset, with the model using the CLS token as aggregation method, and using as inputs 10-second segments. The displayed confusion matrices show the aggregated results from the 10 folds, with their rows normalized. From that figure, we notice that the model is better at recognizing the high category than the low category of arousal. For example, when using ECG signals to predict arousal (Figure 3.9a), the model identifies correctly 91% of the high arousal samples, compared to 82 % of the low arousal samples. When recognizing valence from ECG (Figure 3.9c), the model obtains better results when identifying the low category than the high category. Meanwhile, when recognizing valence from EEG (Figure 3.9d), the model has equal performance when recognizing the low and the high category. In summary, our model is better at recognizing negative emotions of high intensity. We found this result satisfactory, in the sense that we believe that this is the most critical situation that should be identified when monitoring the emotional state of a frail person.

	Pre-train	Arousal Acc.	Arousal F1	Valence Acc.	Valence F1	Avg. Δ
AMIGOS ECG	No	0.85±5.6e^{-3}	0.84±5.8e^{-3}	0.80±6.5e^{-3}	0.80±6.4e^{-3}	3.7%
	Yes	**0.88**±5.4e^{-3}	**0.87**±5.4e^{-3}	**0.83**±7.8e^{-3}	**0.83**±7.4e^{-3}	
DREAMER ECG	No	0.74±1.1e^{-2}	0.74±1.2e^{-2}	0.72±8.2e^{-3}	0.71±7.4e^{-3}	11.7%
	Yes	**0.83**±7.1e^{-3}	**0.83**±7.6e^{-3}	**0.80**±1.1e^{-2}	**0.79**±1.1e^{-2}	
AMIGOS EEG	No	0.76±7.3e^{-3}	0.75±8.3e^{-3}	0.70±6.4e^{-3}	0.70±6.8e^{-3}	8.3%
	Yes	**0.81**±1.1e^{-2}	**0.80**±9.4e^{-3}	**0.77**±9.3e^{-3}	**0.77**±9.1e^{-3}	
DREAMER EEG	No	0.64±1.0e^{-2}	0.64±1.1e^{-2}	0.63±1.1e^{-2}	0.61±1.1e^{-2}	7.8%
	Yes	**0.68**±1.7e^{-2}	**0.68**±1.6e^{-2}	**0.68**±1.9e^{-2}	**0.67**±1.4e^{-2}	

Table 3.4 – No Pre-trained vs. pre-trained model for the different evaluation scenarios. Avg. Δ is the average percentage increase of the different metrics between the no pre-trained model and its pre-trained counterpart.

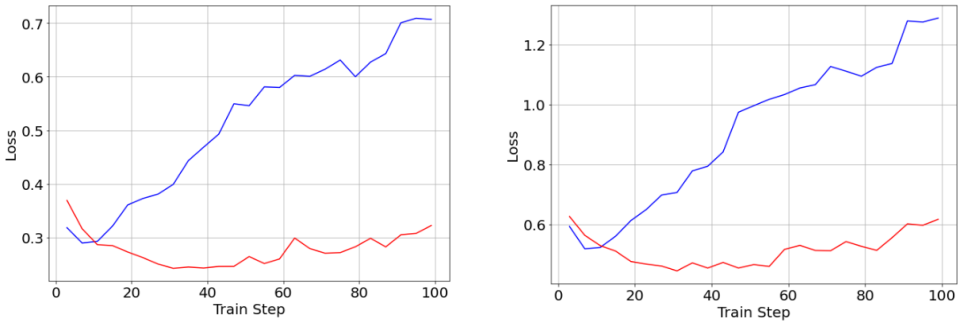
3.4.2.3 Effectiveness of Pre-Training

To evaluate the effectiveness of our pre-training strategy and validate Expected Result 3.1, we replace our pre-trained signal encoder with an encoder using randomly initialized parameters. In other words, we skip the left part of our process depicted in Figure 3.6.

As can be seen from both the accuracy and the F1-score shown in Table 3.4, a pre-trained model was found to perform better than a model trained from scratch in all the evaluation scenarios. These results demonstrate that pre-training our Transformer-based signal encoder is beneficial for our task of recognizing emotions from physiological signals. Pre-training allows the model to build stronger representations, making the model more generalizable and less prone to overfitting. These ideas are expanded below

A more generalizable model: In Table 3.4, the column Avg. Δ indicates the average percentage increase between a pre-trained and a no pre-trained model. The average is calculated across all the metrics of each evaluation scenario. We can see that the minimum increase is 3.7%, and it corresponds to using AMIGOS with ECG signals. For the other evaluation scenarios, the increase is significantly higher.

In developing the architecture, we used the ECG signals from the AMIGOS dataset for both tuning and testing the model. As a result, the hyper-parameters of the model are better tuned for this scenario than for the other ones. Therefore, even without pre-training, in this case the model reaches high performance, so improving is harder. In the other scenarios, the



(a) Validation losses for Arousal (b) Validation losses for Valence

Figure 3.10 – Comparison of the losses on the validation set when using a pre-trained model, in red, compared to using a model without pre-training, in blue. We show the losses for arousal (a) and valence (b).

architecture may not be completely tuned for them, so we get relatively low scores when no pre-training is used. But using pre-training, we overcome the problem of not using a highly tuned model, gaining large increases in performance. From these observations, we can conclude that through our pre-training strategy we make the model more generalizable, allowing it to perform well across different datasets and physiological signals.

Overcoming overfitting: During our experiments with pre-trained and not pre-trained models, we noticed that the model without pre-training had a tendency to overfit quickly, while the pre-trained model did not exhibit the same behavior. Figure 3.10 shows an example of this, where we compare the losses in the validation dataset between using a pre-trained (red line) and a no pre-trained (blue line) model. This figure shows the average validation loss across the 10 folds for the AMIGOS dataset with ECG signals. We obtained similar results in the other evaluation scenarios. Through these observations, we can conclude that pre-training the model on different datasets increases its robustness to overfitting when the model is fine-tuned on a specific dataset.

3.4.2.4 Comparison With Other Approaches

Table 3.5 reports various state-of-the-art results for emotion recognition obtained using the same datasets and the same physiological signals that we use. It is necessary to take into account that these works use different experimental protocols to perform the evaluation. For example, there is a variety of input segment sizes, different partitions of data into training and test sets, subject-dependent and independent evaluations, etc. For this

Model	Subj. Ind.	Input Seg. Size	Arousal Acc.	Arousal F1	Valence Acc.	Valence F1
AMIGOS WITH ECG SIGNALS.						
Gaussian Naive Bayes [135]	Yes	20s	-	0.55	-	0.55
1D-CNN [161]	No	200 peaks	0.81	0.76	0.71	0.68
2D-CNN [175]	Yes	No	0.83	0.76	0.82	0.80
1D-CNN with LSTM [85]	Yes	No	-	-	0.81	0.80
Pre-trained CNN [162]	No	10s	0.89	0.88	0.88	0.87
Autoencoder [154]	No	10s	0.85	0.89	-	-
Pre-trained Transf. (ours)	No	10s	0.88	0.87	0.83	0.83
DREAMER WITH ECG SIGNALS.						
SVM [102]	Yes	No	0.62	0.58	0.62	0.53
2D-CNN [175]	Yes	No	0.81	0.77	0.80	0.78
1D-CNN with LSTM [85]	Yes	No	-	-	0.71	0.66
Pre-trained Transf. (ours)	No	10s	0.83	0.83	0.80	0.79
AMIGOS WITH EEG SIGNALS.						
Gaussian Naive Bayes [135]	Yes	20s	-	0.58	-	0.56
2D-CNN [175]	Yes	No	0.79	0.74	0.83	0.80
CNN + SVM [187]	No	No	0.91	-	0.87	-
1D-CNN [114]	Yes	No	0.66	0.67	0.61	0.63
2D-CNN [114]	Yes	No	0.79	0.79	0.79	0.76
Pre-trained Transf. (ours)	No	10s	0.81	0.80	0.77	0.77
DREAMER WITH EEG SIGNALS.						
SVM [102]	Yes	No	0.62	0.58	0.62	0.52
2D-CNN [175]	Yes	No	0.79	0.77	0.79	0.75
Graph CNN [178]	No	60s	0.85	-	0.86	-
CNN + SVM [187]	No	No	0.90	-	0.88	-
1D-CNN [114]	Yes	No	0.61	0.63	0.61	0.60
2D-CNN [114]	Yes	No	0.83	0.81	0.80	0.79
Pre-trained Transf. (ours)	No	10s	0.68	0.68	0.68	0.67

Table 3.5 – Results of different methods on the different evaluation scenarios. These results are not directly comparable as the experimental protocols are not necessarily the same.

reason, we cannot compare these results directly to each other, nor can we compare these results with our work. However, we present them to demonstrate the range of different solutions that have been proposed for this task and to provide a comparative understanding of the performances achieved in the different evaluation scenarios.

To have a more fair comparison of our work and another state-of-the-art approach, it is necessary that both use the same experimental protocol. In order to achieve this, we fully retrain and evaluate the pre-trained CNN approach proposed by Sarkar and Etemad [162], using exactly the same

Model	Arousal Acc.	Arousal F1	Valence Acc.	Valence F1
AMIGOS WITH ECG SIGNALS.				
Pre-trained CNN [162]	0.85±5.4e^{-3}	0.84±5.3e^{-3}	0.77±5.5e^{-3}	0.77±5.1e^{-3}
Pre-trained Transf. (ours)	**0.88**±5.4e^{-3}†	**0.87**±5.4e^{-3}†	**0.83**±7.8e^{-3}†	**0.83**±7.4e^{-3}†
DREAMER WITH ECG SIGNALS.				
Pre-trained CNN [162]	0.81±1.1e^{-2}	0.81±9.9e^{-3}	0.79±8.4e^{-3}	0.78±7.3e^{-3}
Pre-trained Transf. (ours)	**0.83**±7.1e^{-3}†	**0.83**±7.6e^{-3}†	**0.80**±1.1e^{-2}	**0.79**±1.1e^{-3}†

Table 3.6 – Comparison of our approach with the approach of Sarkar and Etemad [162], under the same experimental protocol. The symbol (†) indicates that the differences are statistically significant.

protocol that we use in our experiments. To run these experiments, we use the code provided by the authors [1]. The work of Sarkar and Etemad addresses the task of recognizing emotions from ECG signals, so we make the comparison only with the evaluation scenarios where ECG signals are used. To pre-train, train and test their approach, we use the same data that we use to pre-train and train our model. We also use the same partitions of train, validation, and test sets, and the same folds when doing the 10-fold cross-validation. In general, where applicable, we replicate the experimental setup described in Section 3.4.1.

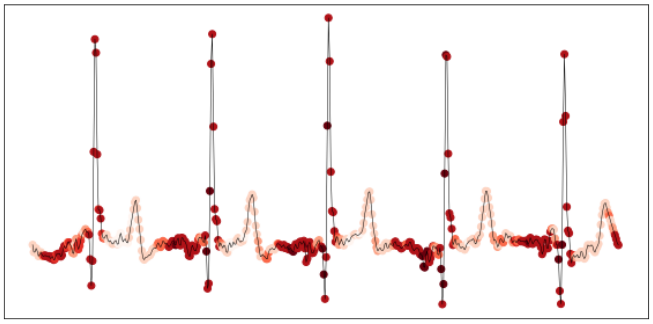
Table 3.6 shows that under the same experimental protocol, our approach achieves better performance than the approach of Sarkar and Etemad, for both arousal and valence. Moreover, the difference between the results of our and their approach is statistically significant, with $p < 0.05$ following a t-test. This last statement is true for all the results except for valence accuracy using the DREAMER dataset. In this case, although our results are better, they are not statistically different than those obtained using their approach. These results confirm our Expected Result 3.2, i.e. we expected that our approach should improve the state-of-the-art.

With these results, we can conclude that our pre-trained Transformer approach produces strong representations useful for predicting emotions from ECG and EEG signals, which improves the results over the previous state-of-the-art.

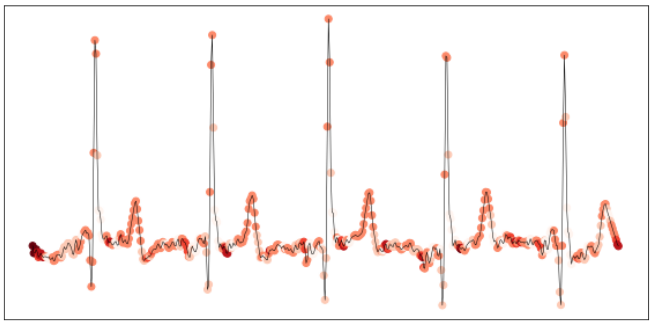
3.4.2.5 Attention Weights

It is interesting to observe the attention weights produced by the Transformer, to assert our claim that the Transformer is capable of assigning more

1. https://code.engineering.queensu.ca/pritam/SSL-ECG



(a) Attention weights when predicting arousal.



(b) Attention weights when predicting valence.

Figure 3.11 – Attention weights overlaid in the corresponding ECG input signal, corresponding to arousal prediction (a) and valence prediction (b). The darker the color, the greater the attention weight.

importance to certain parts of the signal, i.e. checking if our Expected Result 3.3 is correct. To do this, we extract the Transformer attention matrix from the last layer of the Transformer. Since in our experiments each layer has two heads, we calculate the average of the weights produced by each head.

Figure 3.11 shows the attention weights overlaid over the corresponding ECG input signal of a chosen sample. The darker the color, the greater the attention weight. These weights correspond to the CLS token; that is, it indicates how much attention this token pays to each point on the input signal. Recall that this token is used to perform the classification.

It is interesting to see that the model succeeds in capturing the periodic nature of the ECG signal. For example, Figure 3.11a clearly shows that the attention weights follow a periodic pattern, each time giving less attention to the small peaks that come after the main peak.

These observations confirm our claim that a Transformer based approach is capable of assigning different weights to the input signal, thus allowing us to obtain a representation that is the weighted aggregation of the information from all the signal, confirming this way our Expected Result 3.3.

3.5 Conclusions

This chapter presented the first contribution of this thesis: pre-training a Transformer with a self-supervised approach for emotion recognition from physiological signals. With this contribution, we address the two main challenges that were identified regarding this task: the lack of large quantities of labeled data to train the model (Challenge 3.1), and how to process raw physiological signals effectively (Challenge 3.2).

First, to address the challenge of how to process raw physiological signals effectively, we employ a Transformer-based approach. We showed that in fact this architecture assigns more weight to certain parts of the input signal, which we could argue are the *important* parts. Second, to address the challenge of not having large quantities of labeled data, we pre-train our model with unlabeled physiological signals using a self-supervised approach. We experimentally demonstrated that our pre-training technique, predicting masked segments of the inputs, leads to a performance improvement compared to using a no pre-trained model. In addition, we discussed that this improvement could be explained because the model becomes more generalizable and less prone to overfitting. Talking more broadly, with the results presented in this chapter, we showed that self-supervised pre-training and Transformer-based models could be successfully used in affective computing.

Taking a wider view, if the goal is to monitor the mental well-being of frail people in a smart environment, it is important to recognize emotions accurately, using signals that such an environment may provide. This environment can be equipped with devices that collect signals for other health purposes, like physiological signals. Thus, it is appealing to use that type of signal for emotion recognition. Moreover, nowadays those signals can be acquired using portable, affordable off-the-shelf devices, therefore in the future, we expect that the acquisition of those signals becomes even less invasive. We believe that the contributions made in this chapter are a step towards the goal of monitoring the emotional wellness of frail people living in smart environments.

In this chapter, we used each type of physiological signal independently and obtained good results with each signal. This shows that those signals contain useful information for emotion recognition. It may be the case that this information may not overlap. Therefore, if multiple physiological signals are used simultaneously, they may complement each other, thus further improving the results. This approach is explored in the next chapter, where we investigate the usage of multiple physiological signals simultaneously to perform emotion recognition.

CHAPTER 4

EMOTION RECOGNITION FROM MULTIPLE PHYSIOLOGICAL SIGNALS

To effectively monitor the emotional well-being of a frail person living in a smart environment, it is desirable to have an accurate emotion recognition system. Therefore, it is appealing to take advantage of the multiple physiological signals that may be gathered in a smart environment, not only processing them individually, as it was done in Chapter 3, but using them at the same time to exploit the complementary information that they might carry.

This chapter describes a method to recognize emotions from multiple physiological signals, extending the work from Chapter 3 where we developed a solution that works for single physiological signals. This chapter starts by providing in Section 4.1 a definition of the problem, and a description of the motivations and challenges of this problem. It continues by reviewing the current literature and techniques related to recognizing emotions from multiple sources in Section 4.2. Next, in Section 4.3, it describes in detail the second contribution of this thesis, which is a method

to recognize emotions from multiple physiological signals using pre-trained Transformers. Finally, in Section 4.4, it shows the experimental results obtained when testing our approach.

4.1 Problem Definition, Motivations and Challenges

In this chapter, our goal is to recognize high and low categories of arousal and valence from multiple raw physiological signals. Specifically, raw Electrocardiogram (ECG) and Electroencephalogram (EEG) signals are combined. These are the same signals used in Chapter 3, where they were used individually as inputs for an emotion recognition system. As for chapter 3, those signals are selected because of advantages such as that they can be gathered with portable equipment, they might be already acquired from frail people for other medical purposes, and there exist abundant unlabeled datasets with these types of signals that can be used to pre-train a model.

4.1.1 Motivations

The problem addressed in this chapter is related to the one addressed in Chapter 3, sharing the motivations behind working with high/low categories of arousal and valence, and using raw physiological signals. These motivations, which are explained in detail in Sections 3.1.1.1 and 3.1.1.2, are summarized below:

— **Using high/low categories of arousal and valence:** A way to monitor the mental well-being of a person is to identify if the person is feeling a positive or negative emotion and if the intensity of this emotion is high or low. To identify these situations, it is enough to recognize high and low categories of arousal and valence.

— **Using raw physiological signals:** ECG and EEG signals can be captured with wearable, wireless, and low-cost off-the-shelf equipment, giving the potential to use emotion recognition methods in everyday scenarios [102]. In addition, raw signals are used instead of engineered features extracted from these signals because a Deep Learning (DL) approach is employed, which is capable of extracting robust features from these raw signals.

Multiple physiological signals are expected to improve the accuracy of estimates of the emotional state of a subject by providing complementary

information. Therefore, it is worth trying to combine them in an attempt to improve the accuracy of the results given by the model. Specifically for ECG and EEG signals, they are produced by different systems of the body: ECG signals are related to the autonomic nervous system, and EEG signals monitor brain activity so they are more related to cognition. For this reason, the interaction of each of those signals with emotions is not the same, and therefore, the information about emotions that they carry might be different but complementary.

Since we consider that all physiological signals are from the same modality, we designate this problem as detecting emotions from *multi-signal* inputs rather than from *multimodal* inputs. Nevertheless, both types of problems are related, and the multi-signal emotion recognition task can be seen as a special case of multimodal emotion recognition.

4.1.2 Challenges

In solving the problem addressed in this chapter, there are challenges similar to the ones identified in Chapter 3 and detailed in Section 3.1.2, namely, how to process raw signals effectively and the lack of large quantities of labeled data to train the model.

The challenge of the lack of large quantities of labeled data is especially important in this chapter because there is the need for labeled datasets that have all the concerned signals, which may limit even more the available datasets. With this, the main challenge for solving the problem addressed in this chapter is:

Challenge 4.1. *How to train a model for emotion recognition from multiple physiological signals, without large quantities of labeled data that contain all those multiple physiological signals.*

4.2 State of the Art on Emotion Recognition from Multiple Signals

This section reviews relevant literature related to recognizing emotions from multiple physiological signals, discussing fusion techniques in Section 4.2.1, and then reviewing contributions that use Machine Learning (ML) approaches with and without pre-training in Sections 4.2.2 and 4.2.3.

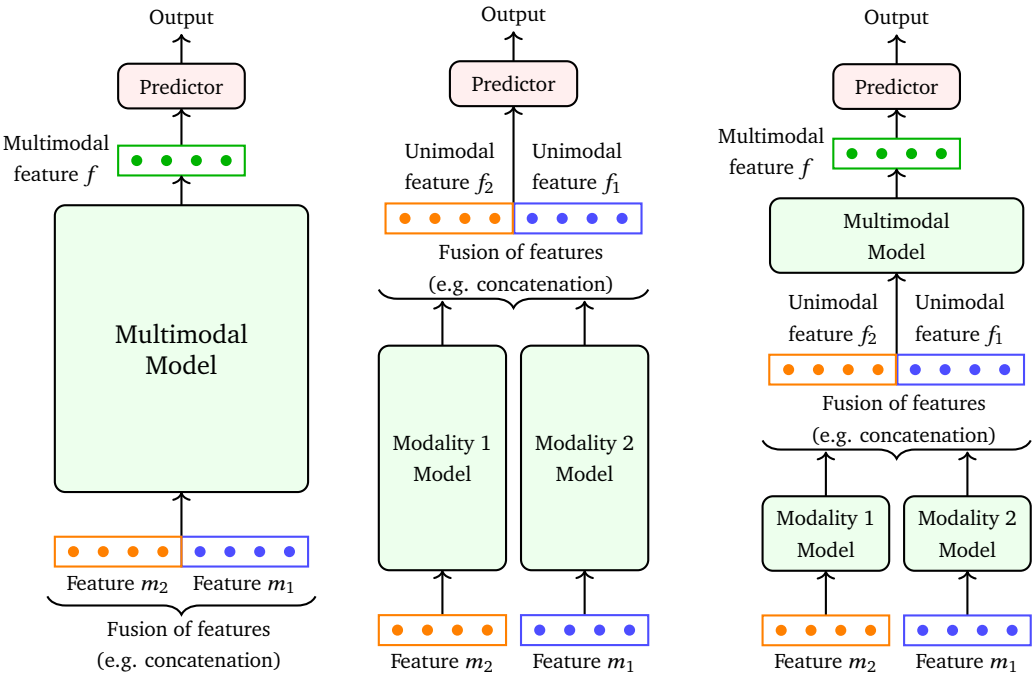


Figure 4.1 – Different types of multimodal fusion.

4.2.1 Fusion of Multiple Signals

It is possible to use established techniques of multimodal fusion to combine multiple physiological signals. As described by Baltrusaitis et al. [17], most of the works that employ multimodal fusion use a model-agnostic approach that can be divided into two types: early fusion and late fusion. If we constrain the models to DL approaches, a third type of fusion can be introduced: mid-fusion. The difference between those approaches is at what level the fusion of the different modalities takes place. Figure 4.1 shows a graphical representation of those types of fusion.

4.2.1.1 Early Fusion

As described in Atrey et al. [10], early fusion is done by combining the different modalities before they are fed into the model. Commonly, this is done by concatenating features extracted from each modality, although raw data can also be used as input. This approach is depicted in Figure 4.1a, where m_1 and m_2 are the features from two different modalities. This type of fusion allows the model to learn cross-correlations and interactions between

the low-level characteristics of each modality [74]. In fact, with early fusion, we can say that the obtained feature f is a multimodal representation of the input signals.

One of the advantages of early fusion is that we only work with a single model. This means that the training process is typically less cumbersome than mid- and late-fusion approaches.

4.2.1.2 Late Fusion

As described in Gadzicki et al. [74], late fusion combines the different modalities after each one has been processed independently. One way of doing this is to obtain the predictions (decisions) for each modality and then combine those decisions to obtain the final one [10]. For this reason, this approach is often called decision-level fusion. As described in Cumin and Lefebvre [46], the final decision can be obtained through mechanisms like, voting, or stacking. In voting methods, the decision obtained by each modality is used as a vote, and a fuser module collects those votes and produces the final result. In stacking methods, a top-level model is trained to use as input the decisions of each modality and produce the final result.

When using a DL model, another option is that instead of using the individual decisions, a late-fusion model can use the features produced by each individual model before the decisions are made. Then, those features can be combined by training a new model that uses them as input. Figure 4.1b depicts this process.

In any case, late fusion combines the outputs of two or more uni-modal models. Commonly, each individual model is trained independently, although all of them are trained for the same task. As pointed out by several authors like Atrey et al. [10], in a late-fusion approach each model can specialize in processing its corresponding modality, obtaining better uni-modal features, but on the other hand, the low-level relations between the modalities are ignored.

4.2.1.3 Mid-Fusion

Instead of fusing the modalities at the beginning, like in early fusion, or at the end, like in late fusion, a middle-ground approach is possible, as described by authors like Liu et al. [125] and Nagrani et al. [139]. The idea of this approach, called mid-fusion, is to process modalities individually, and then the obtained representations are combined and further processed

together. This allows the first layers of the architecture to model the low-level individual characteristics of the signal, and the upper layers model inter-modality relations. Figure 4.1c shows a depiction of this approach.

4.2.1.4 Discussion About Fusion Approaches

In this chapter, we employ late fusion to combine multiple physiological signals to perform emotion recognition because, as it will be described in Section 4.3, pre-training is used in our approach, and using early fusion or mid-fusion may impose restrictions on the datasets used for pre-training the model. With pre-training, the idea is to use many different datasets in order to obtain a more robust representation of the information of the different signals. This collection of datasets does not need to be related to the task of emotion recognition. If early fusion is employed, the pre-training datasets should include all the targeted types of signals. This severely limits the availability of datasets that could be used. Conversely, if late fusion is used, each uni-signal model can be pre-trained independently, thus having the possibility of using different collections of datasets, each one including only the concerned signal.

Our late-fusion approach uses a top-level model that uses as input the concatenated features of individual models. Other late-fusion aggregation methods might be considered, like using the decisions of individual models for voting or stacking. However, we believe that a top-level model using the features produced by individual models will help better aggregate the complementary information from each signal.

4.2.2 Machine Learning Approaches for Emotion Recognition from Multiple Physiological Signals

Classical ML techniques such as Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), and Decision Trees have been used by several authors [80, 102, 135] to perform emotion recognition from multiple physiological signals. Typically, these techniques use as inputs engineered features extracted from the signal.

On the other hand, DL architectures extract data-driven features that typically lead to more accurate results than using engineered features, as demonstrated in works like Santamaria-Granados et al. [161] and Siddharth et al. [175]. In addition, DL approaches allow the usage of self-supervised pre-training techniques, making it possible to take advantage of unlabeled

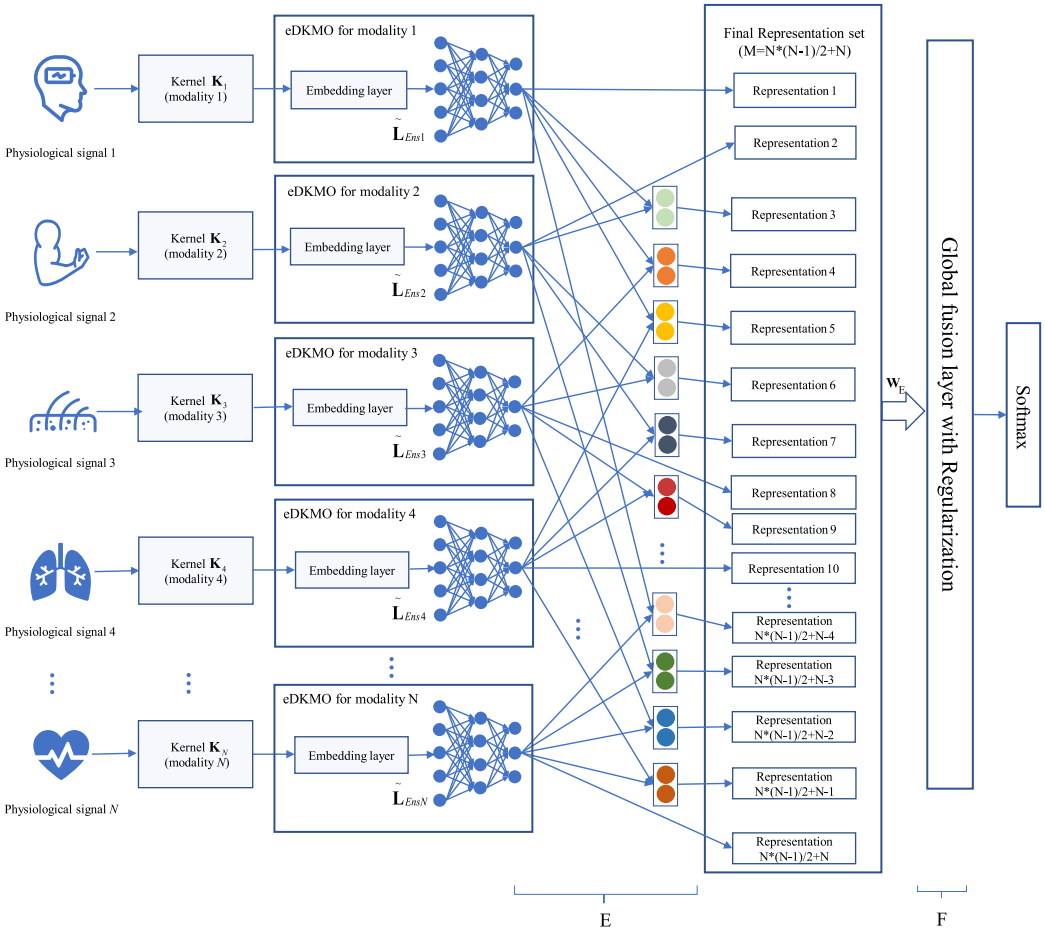


Figure 4.2 – Zhang et al. [221] approach. (Figure from Zhang et al.'s paper [221]).

datasets to further improve the accuracy of the results compared to a fully-supervised approach.

Below, we provide a review of some contributions that use ML approaches to perform emotion recognition, focusing on approaches that employ DL techniques.

4.2.2.1 Emotion Recognition with Deep Fusion of Kernel Machine

Zhang et al. [221] use kernel matrices to construct an ensemble of dense embeddings from each input signal and then process these embeddings using a Fully-Connected Network (FCN) to obtain the signal representations. To combine the different signals, they use a global fusion layer.

In Zhang et al.'s approach [221], which is depicted in Figure 4.2, the

authors use as inputs engineered features extracted from the physiological signals. The physiological signals that they use are EEG, Electromyogram (EMG), Electrodermal Activity (EDA), and Respiration (RESP). The first step is to obtain embeddings for each type of signal using a kernel matrix constructed with a kernel function. Specifically, several mapping functions are built from the kernel matrix using different sample subsets. This is done because having multiple mapping functions allows the modeling of the characteristics of different regions in the input space. From these embeddings, a representation of each type of signal is obtained using a multilayer FCN.

To combine the representations from the signals, an intermediate fusion step is performed first. In this step, several one-layer FCN are used to do a pair-wise combination of the representations of the different signals. Next, a global fusion layer is used to obtain the final fused representation. The inputs for the global fusion layer are the representations of each type of signal, plus the outputs from the intermediate fusion step. The predicted class distributions are obtained using a soft-max function on the final fused representation.

The work of Zhang et al. [221] shows that engineered features can be processed using a DL to learn better suitable representations for emotion recognition. Moreover, their work exemplifies the usage of a late-fusion approach by first obtaining representations of each physiological signal, and then combining this signal with a fusion layer that works as the predictor. Another important takeaway from this work is that although they obtain the best result when combining all four physiological signals, combining two or three signals does not necessarily outperform using a single signal.

4.2.2.2 Using Features From Pre-Trained Deep Learning Models

Siddharth et al. [175]. take advantage of DL models trained for computer vision to extract features for the task of emotion recognition from multiple physiological signals. An Extreme Learning Machine [92] is used to process these features to predict emotions.

Siddharth et al. [175] use a VGG model [176] to extract features from EEG, ECG, Photoplethysmography (PPG) and EDA signals. VGG is a model trained for object recognition using a dataset with more than one million images and 1000 object classes. In order to use VGG to process physiological signals, the signals need to be transformed into images. To do this for EEG signals, they extract the power spectral density of the signal, and then they

produce a heat map using these values plotted and interpolated in a 2D plane according to the location of the EEG electrodes. For ECG, PPG, and EDA signals, they generate spectrogram images [71] from each of those signals, which are resized to fit in the VGG model. The VGG model produces features of size 4096, and the authors use Principal Component Analysis (PCA) to reduce this dimension to 30. The features obtained with the VGG model are concatenated and then processed with an Extreme Learning Machine [92] to predict high and low categories of arousal and valence.

This work shows that DL models are capable of producing data-driven features for emotion recognition. Moreover, in their paper, Siddharth et al. [175] visually compare the feature spaces obtained with engineered features with the feature spaces from the VGG features and show that the latter allows for better separation of classes. When comparing the results of using multiple signals with using single signals, the authors found that in the majority of experiments they conducted on different datasets, using multiple signals outperforms using single signals.

4.2.2.3 Emotion Recognition with Classical Machine Learning Approaches

Miranda-Correa et al. [135] introduce the AMIGOS dataset and use a classical ML model to perform emotion recognition on that dataset. This model uses as inputs engineered features extracted from ECG, EEG, and EDA signals. For each signal, the authors train a GNB classifier, and a decision-level fusion is implemented to combine the outputs from the different signals, using a SVM as the predictor. Interestingly in their case, most of their experiments show better results when using only EEG rather than combining the three signals.

Katsigiannis and Ramzan [102] introduce the DREAMER dataset and perform emotion classification with a SVM employing a radial basis function kernel, using as inputs engineered features extracted from ECG and EEG signals. In this work, early fusion is used to combine the signals, concatenating the features from ECG and EEG, and feeding the concatenated vector into the SVM. The results in Katsigiannis and Ramzan's [102] paper show that the difference in performance between fusing the signals and using the signals independently is not statistically significant, showing that the engineered features from the different signals describe the same information about the emotional state of the person.

4.2.2.4 Discussion on Emotion Recognition from Multiple Physiological Signals

Several conclusions can be made from the above review. First, concerning the use of DL-based approaches, Zhang et al. [221] and Siddharth et al. [175] show that for the problem of emotion recognition from multiple signals, extracting features with DL leads to better results than using engineered features with a non DL model. This is in line with other works that use a single physiological signal (see Section 3.2.2).

Second, the reviewed works show that combining multiple signals may lead to an improvement in performance compared to using single signals, even though sometimes it may be difficult to know beforehand if fusing the signals will increase performance or which signals to combine.

Finally, although the reviewed works show the strength of DL approaches, there is an advantage that was not exploited by them, which is using pre-training techniques. Only Siddharth et al. [175] use a pre-trained VGG model, but this model was pre-trained for computer vision tasks, and it could be more convenient to pre-train a model for signal processing. Pre-training can be especially useful when labeled data is scarce, as is usually the case for data with labels of emotion. This data scarcity is the reason why we identified Challenge 4.1. Therefore, in the next section, we review contributions that use pre-training in their approach for the task of emotion recognition from multiple physiological signals.

4.2.3 Pre-trained Models for Multi-Signal Emotion Recognition

As discussed in Chapter 3, pre-training may help the model obtain better results. Many authors have explored this technique when developing approaches for emotion recognition. Several works employ pre-training when using other modalities besides physiological data, like images, sound, and text. Recent works are based on Transformers, like the work of Khare et al. [105], which uses Transformers that are pre-trained by masking some words in the input text, along with the audio and visual parts that correspond to those words, and then the pre-training task consists in predicting the masked words.

Regarding using multiple physiological signals, a common pre-training approach is to use autoencoders to extract representations from the inputs. An autoencoder is an architecture that can learn to extract representations

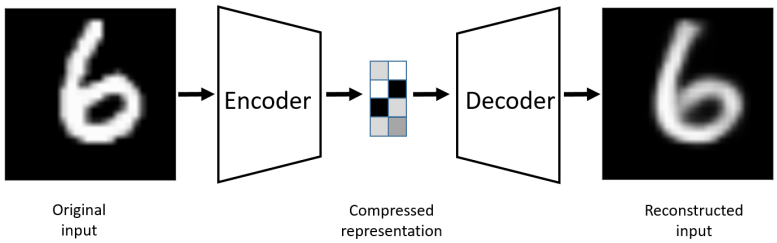


Figure 4.3 – An autoencoder example. (Figure from Bank et al. [18]).

in a self-supervised way. This is done by mapping the input to a latent representation and then reconstructing the input from this representation, as depicted in Figure 4.3. The pre-training can be done by comparing the reconstructed input with the original input. The remainder of this section reviews some papers that exemplify this approach.

4.2.3.1 Extracting Representations with Autoencoders

Ross et al. [154], report on the use of Variational Autoencoders (VAEs) to extract representations from raw ECG and EDA signals, for the task of predicting arousal. They train two independent VAEs, one for each signal. The VAEs are based on stacked layers of 1D Convolutional Neural Networks (1D-CNNs). The combination of the representations learned by the ECG and EDA autoencoders is done by concatenating those representations. Then, a Random Forest model is used to predict arousal using as inputs the concatenated representations.

Ross et al. [154] compare using VAE representations to using engineered features and show that the former approach improves the accuracy of the results. In addition, they also show that fusing the two signals has better performance in terms of accuracy and F1-score than using the signals individually. In summary, this work exemplifies that pre-training techniques can be used to extract relevant representations from the signals for the task of emotion recognition from multiple physiological signals.

4.2.3.2 Emotion Recognition with a Bimodal Autoencoder

Liu et al. [127] address the task of predicting positive, neutral, and negative emotions from EEG signals and eye movement or other peripheral physiological signals, depending on the dataset. First, they train two independent Restricted Boltzmann Machines (RBMs), one for each signal, using as inputs engineered features extracted from the signals. Once the RBMs

are trained, the outputs from those machines are concatenated and used as input for an upper RBM. The output for the upper RBM is the shared representation of the multiple signals, which is used as input to reconstruct the signals during the unsupervised pre-training phase. This reconstruction is done using a network with the same weights as the encoding RBMs. The shared representations generated by the pre-trained model are used to train an SVM classifier.

When the authors test their approach on a first dataset, they show that using the shared representations obtained from EEG and eye movement leads to better results than a uni-signal model. Moreover, they show that simply using the concatenation of the engineered features of both signals directly as input for the SVM is better than using the signals independently.

When testing on a second set using EEG and peripheral physiological signals as input, the authors also find that their pre-trained multi-signal approach improves over using uni-signal models. However, in this case, if the concatenation of engineered features is used instead of the shared representation, there is a decrease in performance compared to the uni-signal models.

The results from Liu et al. [127] show that simply combining different modalities may not lead to an improvement in performance, but this improvement can be achieved if robust representations are used to do the emotion recognition. Moreover, they show the usefulness of pre-training in improving the results when doing emotion recognition from multiple physiological signals.

4.2.3.3 Discussion on Pre-Trained Models

The reviewed literature shows that it is possible to use pre-training techniques to extract robust representations for multi-signal emotion recognition. Moreover, Ross et al. [154] and Liu et al. [127] show that these representations can better model the complementary information that may be present in the different physiological signals, thus making the results better when using multiple signals than when single signals are used.

4.2.4 Discussion

Several conclusions can be obtained from the contributions studied in this section. First, DL are capable of extracting features useful for the task of multi-signal emotion recognition that improves the results over using

engineered features. Moreover, DL approaches have the advantage that can be pre-trained using self-supervised approaches.

Second, using multiple signals may help to improve the performance compared to using single signals. This depends in part if the features used are capable of modeling the complementary information that may be present in the signals, and depends also on using a model capable of exploiting this complementarity, if it exists.

Third, pre-trained models help to obtain representations that improve the accuracy of the results produced by the model, compared to no pre-trained approaches. Moreover, these representations help the model to better take advantage of using multiple physiological signals.

Following these conclusions, we aim to develop a DL architecture that uses pre-training as part of the training process, to perform emotion recognition using multiple raw physiological signals. Specifically, our contribution uses a Transformer-based approach to process raw physiological signals because, as demonstrated in Chapter 3, this architecture is capable of processing raw physiological signals effectively. In addition, we use pre-training to overcome the issue of not having large labeled datasets that include all the concerned physiological signals, as pointed out by Challenge 4.1. To do this, we use a late-fusion approach, that allows us to pre-train individual models using unlabeled datasets that do not necessarily contain all the concerned signals.

Reviewing the literature, we concluded that multimodal pre-training approaches are not typically used on physiological signals, and conversely, pre-training approaches for physiological signals are usually single-modality, and in addition, the few multimodal pre-trained approaches for physiological signals we surveyed don't use attention-based models. Therefore, our contribution aims to investigate how to exploit the advantages of a pre-trained Transformer-based model to perform emotion recognition from multiple raw physiological signals, something not explored in the current state-of-the-art.

4.3 Pre-Trained Transformers for Multi Physiological Signals

This section describes our contribution to the problem of recognizing high and low categories of arousal and valence from multiple physiological

signals. As we shall see, we use a pre-training strategy to address Challenge 4.1, which is not having large quantities of labeled data containing all the considered physiological signals. For the rest of the chapter, the term emotion recognition is used to refer to the recognition of high and low categories of arousal or valence. As it was done in Chapter 3, we use the same architecture to recognize arousal and valence, but train one model to recognize arousal and another to recognize valence.

The backbone of our architecture is a Transformer [196], which is used to process multiple physiological signals, as we did in Chapter 3 for single signals. As shown in that chapter, Transformers employ a learned attention mechanism to dynamically score the relevance of different parts of an input. In other words, Transformers can aggregate information from signals giving more weight to the more relevant parts.

4.3.1 Type of Fusion

As seen in Section 4.2.1, it is possible to combine the information from the different signals at different levels of the model, namely early, late, or in the middle. Therefore, when designing an architecture to process multiple signals, an important problem is to determine at which level those signals should be combined. In our case, as we are interested in a multi-signal model that employs pre-training techniques, we use a late-fusion approach.

When employing late fusion, each individual uni-signal model can be pre-trained individually, allowing the usage of different collections of unlabeled datasets, each one including only the concerned signal. Thus, the number of potential datasets that can be used for pre-training increases, as they do not need to have *all* the concerned signals.

Another problem with early fusion is that the input for the model is generally the concatenation of the features from the different modalities. In our case, this means that the input would be the concatenation of the physiological signals in the temporal dimension, forming a single and longer sequence. However, it is necessary to take into account that the computational complexity of the Transformer is $O(n^2)$, where n is the input sequence length. This means that a longer sequence will be more computationally expensive to process. On the other hand, with a late-fusion approach, it is possible to train several uni-signal models, training them one by one on less powerful hardware than the one required to train a single and more computationally expensive multi-signal model. Moreover, during the fusion phase, the uni-signal models can be frozen, meaning that we only need to

train the predictor model (see Figure 4.1b). Therefore, with late fusion, for the fusion phase, the model can also be trained with less powerful hardware resources. A similar line of thinking can be used to see the benefits of late fusion compared to mid-fusion in this scenario.

4.3.2 Multi-Signal Emotion Recognition Model

This section presents our strategy for recognizing high and low categories of arousal and valence, using multiple physiological signals, specifically, using ECG and EEG signals. Our procedure is performed in two steps. In the first step, we pre-train and fine-tune two uni-signal models. One model is trained to recognize emotions from ECG signals, and the other one is trained to recognize emotions from EEG signals. In the second step, we use late fusion to combine the outputs of the uni-signal models, training a network to recognize emotions from the combined outputs. More details about the architectures employed in each one of these steps are provided below.

4.3.2.1 Uni-Signal Models

For the uni-signal models, we use our approach proposed in Chapter 3. That is, we use a model based on a Transformer [196]. Since we employ a pre-trained approach, the uni-signal models are trained in two phases. First, we pre-train each model by reconstructing masked values in the input signal. Labeled data is not required for this phase, but only data that include the corresponding physiological signal. Second, we fine-tune each model for emotion recognition. This second step is done in a supervised fashion using labeled data. A graphical depiction of the uni-signal model is depicted in Figure 4.4. More details can be found in Chapter 3.

4.3.2.2 Multi-Signal Emotion Recognition Model

We use late fusion to combine the outputs from the ECG and EEG models, as depicted in Figure 4.5. Instead of using the recognized emotions from each individual model (decision fusion), we prefer to use the output of the last hidden layer (not the output layer) from each uni-signal model. We do this because these features contain more information than the recognized category of arousal or valence, and we believe that using them should lead to a model that produces more accurate results.

To combine individual outputs, we simply concatenate them. Other ways of combining them might be considered, like using max or average pooling.

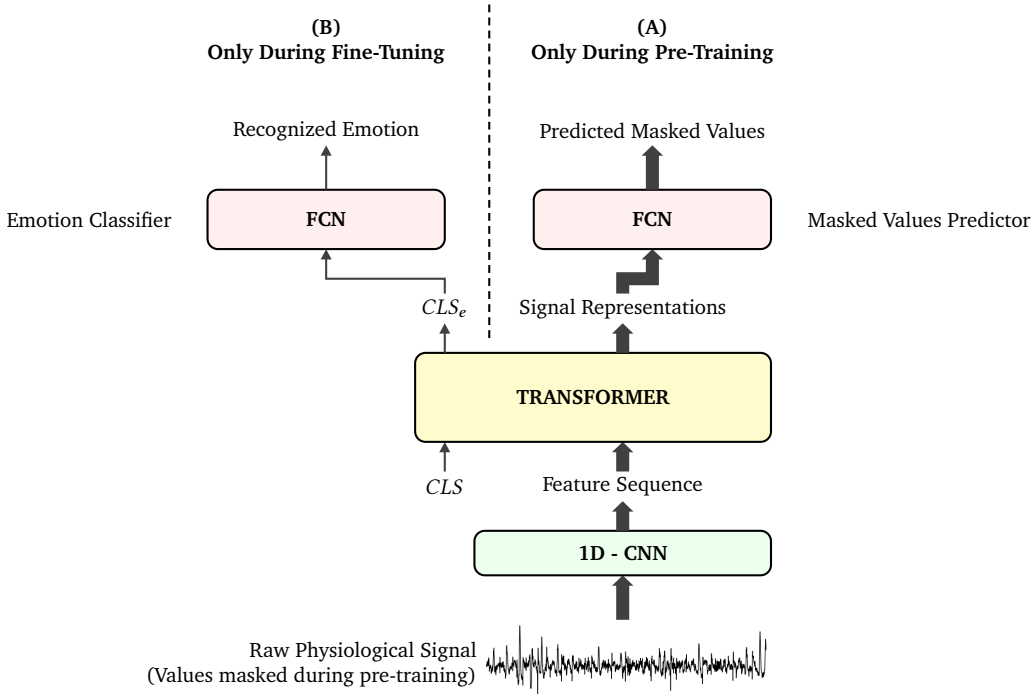


Figure 4.4 – Uni-Signal Model: The raw signal is encoded by a 1D-CNN and processed with a Transformer. First, the model is pre-trained by masking some values of the unlabeled input signal and then predicting those masked values (Part A). Then, labeled data is used to fine-tune the model in a supervised way (Part B).

However, alternative methods may pose some constraints in the design of the models. Pooling, for example, requires that the outputs of each uni-signal model have the same size, which in turn means that the last hidden layers of those models also have the same size. Without this constraint, we can freely choose the size of each individual model that makes the whole approach perform best.

We use a FCN to process the concatenated outputs. This network outputs the predicted high or low category of arousal or valence. That is, the FCN performs emotion classification using the outputs of the individual models. Thus, in Figure 4.5, this FCN network is noted as *Multi-Signal Emotion Classifier*.

When training the fused model, we freeze the weights of both uni-signal models, training only the FCN emotion classifier. As mentioned before, one model is trained to predict high and low categories of arousal, and another model is trained to predict high and low categories of valence.

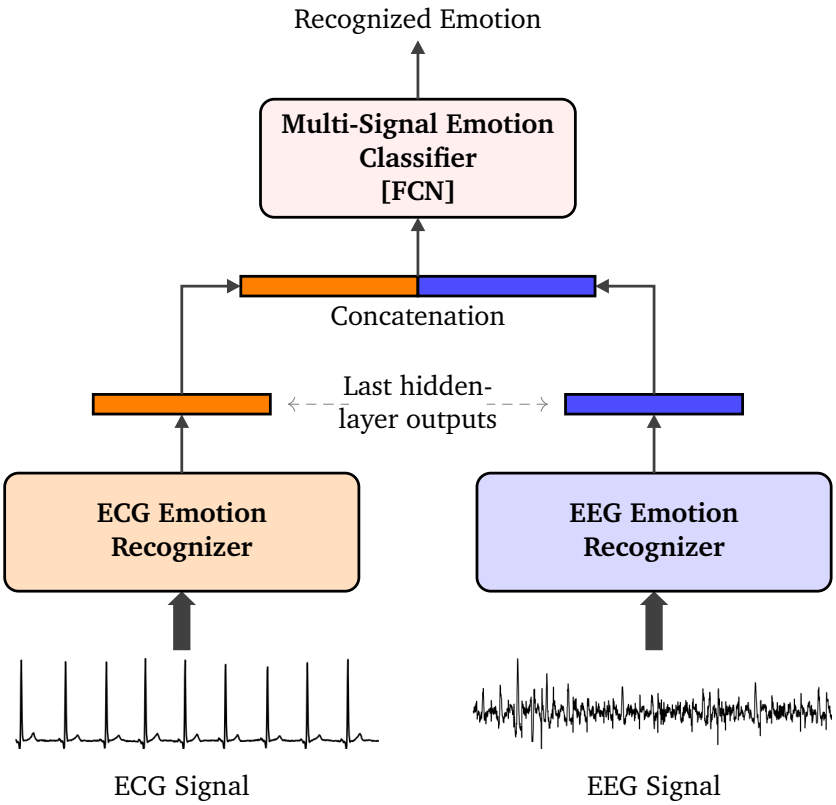


Figure 4.5 – Multi-Signal Model. Late fusion is used to combine the ECG and EEG signals. The outputs of the last layer from both uni-modality models are concatenated, and then used as input to an FCN that outputs the recognized emotion.

4.3.3 Expected Results

4.3.3.1 Expected Results on Using Multiple Physiological Signals

We use multiple physiological signals, specifically ECG and EEG signals, under the assumption that the information that they contain is complimentary. Therefore, if that assumption is true, the following result is expected:

Expected Result 4.1. *Using both ECG and EEG signals at the same time should give more accurate results than using any of those signals individually.*

4.3.3.2 Expected Results on Pre-Training the Model

We use a pre-trained approach in our solution. Specifically, we pre-train and fine-tune individual uni-signal models, and combine them using late fusion. Another option is not to pre-train the individual models, training

each individual uni-signal model from scratch. We refer to this second option as a non-pre-trained model. With this, the following result is expected:

Expected Result 4.2. *Using a pre-trained model should give more accurate results than using a non-pre-trained model.*

4.4 Experiments

This section describes the experimental procedure and the results of testing our approach for emotion recognition from multiple physiological signals, presenting in Section 4.4.1 the experimental setup, giving details about the used datasets and the different hyperparameters of our model, and presenting the results in Section 4.4.2.

4.4.1 Experimental Setup

4.4.1.1 Evaluation Datasets

To train and evaluate our fused-signals model, we use the AMIGOS [135] and the DREAMER [102] datasets, which were also used in Chapter 3. Details about these datasets can be found in Appendix A. We use as labels the self-assessments of arousal and valence provided in those datasets. These assessments are in the ranges 1 to 9 in AMIGOS and 1 to 5 in DREAMER. Since we want to identify high and low categories of arousal and valence, rather than the numerical value, we use the average value in the dataset as the threshold for high and low classes.

Both AMIGOS and DREAMER include ECG and EEG signals. In both cases, for EEG, we use signals taken from the left arm. For EEG, we use the channels F7, F3, T7, P7, O1, O2, P8, T8, F4, F8. These channels are used because they are obtained through electrodes distributed throughout the entirety of the head (see Figure 2.8), thus capturing most of the responses to a given stimulus.

4.4.1.2 Signal pre-processing

We use the same signal pre-processing that we used in Chapter 3 (see Section 3.4.1), filtering the signals with an 8^{th} order Butterworth band-pass filter, with cut-off frequencies of 0.8Hz and 50Hz. Next, the signals are downsampled to a common sample rate of 128Hz. In addition, we normalize the signals so they have zero-mean and unit-variance across each

subject. Finally, we segment the signals into 10-second segments, using each segment as a as sample in our experiments.

4.4.1.3 Uni-Signal Models

We employ the ECG and EEG models described in Chapter 3. That is, we pre-train and fine-tune these models according to the description given in Section 3.4.1. We use the same datasets and the same hyperparameters defined there. Recall that in Chapter 3 a single architecture was developed and pre-trained to recognize arousal and valence, but independent models were fine-tuned, one for arousal and one for valence. Therefore, in this Chapter, there will also be a model to recognize arousal and a different model to recognize valence, although the models will have the same architecture.

4.4.1.4 Multi-Signal Emotion Classifier

Our multi-signal emotion classifier, described in Section 4.3.2.2, is composed of a FCN with two hidden layers, using the Rectified Linear Unit (ReLU) activation function. The sizes of those layers are 64 and 32. We add an output layer to project the result to a single value that corresponds to the predicted binary emotion class.

The network is trained during 52 epochs, with a learning rate of 0.00001 decayed every 20 epochs with a factor of 0.65. We employ a dropout value of 0.1, and Adam optimization with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and L_2 weight decay of 0.00001. The hyperparameters of this network were tuned using the Ray Tune Toolkit [122], with the Bayesian Optimization with Hyperband (BOHB) optimization.

4.4.2 Results

4.4.2.1 Evaluation Strategy and Metrics

To test our approach, we use 10-fold cross-validation across our experiments. We employ the same strategy described in Section 3.4.1 to avoid using the same samples to pre-train and test the model, which in essence is to use two versions of the pre-trained model so that when evaluating with a sample, the evaluation is done with a model that was not pre-trained with that sample.

As metrics, we use accuracy and F1-score, averaged across the two predicted classes (high and low categories of arousal/valence), and averaged

Dataset	Signal	Arousal Acc.	Arousal F1	Valence Acc.	Valence F1
AMIGOS	ECG	$0.88\pm5.4e^{-3}$	$0.87\pm5.4e^{-3}$	$0.83\pm7.8e^{-3}$	$0.83\pm7.4e^{-3}$
	EEG	$0.81\pm1.1e^{-2}$	$0.80\pm9.4e^{-3}$	$0.77\pm9.3e^{-3}$	$0.77\pm9.1e^{-3}$
	ECG+EEG	$\mathbf{0.89}\pm5.0e^{-3}$‡†	$\mathbf{0.89}\pm5.0e^{-3}$‡†	$\mathbf{0.85}\pm3.8e^{-3}$‡†	$\mathbf{0.85}\pm3.9e^{-3}$‡†
DREAMER	ECG	$0.83\pm7.1e^{-3}$	$0.83\pm7.6e^{-3}$	$\mathbf{0.80}\pm1.1e^{-2}$	$\mathbf{0.79}\pm1.1e^{-2}$
	EEG	$0.68\pm1.7e^{-2}$	$0.68\pm1.6e^{-2}$	$0.68\pm1.9e^{-2}$	$0.67\pm1.4e^{-2}$
	ECG+EEG	$\mathbf{0.84}\pm1.3e^{-2}$†	$\mathbf{0.84}\pm1.3e^{-2}$†	$\mathbf{0.80}\pm1.9e^{-2}$†	$\mathbf{0.79}\pm2.2e^{-2}$†

Table 4.1 – Emotion recognition performances of uni-signal models and of the multi-signal model. The symbols (‡) and (†) indicate that the multi-signal result is statistically significantly different than the result with the ECG signal and the EEG signal respectively.

across the 10 folds of cross-validation. We also report the confidence intervals of each metric, calculated across the 10 folds of cross-validation, with a t-distribution with 9 degrees of freedom for a two-sided 95% confidence, using Expression 3.8.

4.4.2.2 Multi-signal Model Results

Table 4.1 shows a comparison of accuracies and F1-scores between the multi-signal model and uni-signal models when recognizing arousal and valence. The uni-signal models used pre-training as part of their training process, and the multi-signal model uses those pre-trained uni-signal models. For the AMIGOS dataset, the multi-signal model performs better than the uni-signal models for both arousal and valence across the metrics that we use. Moreover, when comparing the results of the uni-signal models with the multi-signal strategy for this dataset, the two-tailed P values are less than $1e^{-3}$, thus the difference is extremely statistically significant.

On the other hand, for the DREAMER dataset, there is no statistically significant difference between the performance of using only the ECG signal and using both the ECG and the EEG signal. We believe this is the case because the performance of the EEG model for the DREAMER dataset is particularly lower compared to the ECG model; therefore, the output of the EEG model does possess relevant information that can produce a gain in performance.

The results of this experiment show that using various physiological may lead to an improvement in performance. In fact, in many cases shown in Table 4.1, combining ECG and EEG signals obtains better results than using those signals individually. Moreover, when there is no increment in performance, our multi-signal approach does at least as good as using

Objective	Class	ECG F1	EEG F1	ECG+EEG F1
Arousal	High	0.90	0.84	0.91
	Low	0.83	0.76	0.86
Valence	High	0.81	0.75	0.84
	Low	0.84	0.78	0.87

Table 4.2 – F1-scores of the high and low classes for the ECG, EEG models, and fused models, using the AMIGOS dataset.

a single-signal. With these results, the Expected Result 4.1 is validated, showing that our model is capable of extracting complementary information from each physiological signal, improving the results when the signals are used together.

Now, we further explore the results with the AMIGOS dataset to see if there are characteristics of each ECG and EEG model that show why combining them leads to better results. Concretely, we want to see if one model does better predicting one class and the other does better predicting the other class. For this, Table 4.2 shows the F1-score of the high and low classes individually, instead of averaging them as in the rest of the results shown in this section. In other words, each entry in Table 4.2 gives an idea of how good each model is in distinguishing each class. In the case of arousal recognition, it can be observed that using ECG, EEG, and combining both signals always performs better for the high class. In the case of valence recognition, all the models perform better for the low class. Therefore, for each arousal and valence objective, one uni-signal model is *not* better at predicting a specific class while the other model is better at predicting the other class. Despite this, the table shows that when combining the signals the performance of predicting both classes improves, thus improving the general performance of the model.

4.4.2.3 Effectiveness of Pre-training the Multi-Signal Model

Table 4.3 compares the performances of our multi-signal model when it uses as backbone pre-trained uni-signal models, and when it uses no pre-trained uni-signal models. We can see that when using pre-training, the model achieves better performance compared to not using pre-training, thus confirming Expected Result 4.2. The difference between the pre-trained and no pre-trained approaches have two-tailed P values less than $5e^{-3}$, thus this difference is extremely statistically significant.

Dataset	Pre-train	Arousal Acc.	Arousal F1	Valence Acc.	Valence F1
AMIGOS	No	0.86±4.9e^{-3}	0.85±5.1e^{-3}	0.82±6.5e^{-3}	0.81±6.8e^{-3}
	Yes	**0.89**±5.0e^{-3}	**0.89**±5.0e^{-3}	**0.85**±3.8e^{-3}	**0.85**±3.9e^{-3}
DREAMER	No	0.74±3.1e^{-2}	0.74±3.1e^{-2}	0.72±2.4e^{-2}	0.70±2.4e^{-2}
	Yes	**0.84**±1.3e^{-2}	**0.84**±1.3e^{-2}	**0.80**±1.9e^{-2}	**0.79**±2.2e^{-2}

Table 4.3 – Fused Model: Pre-Training vs No Pre-Training.

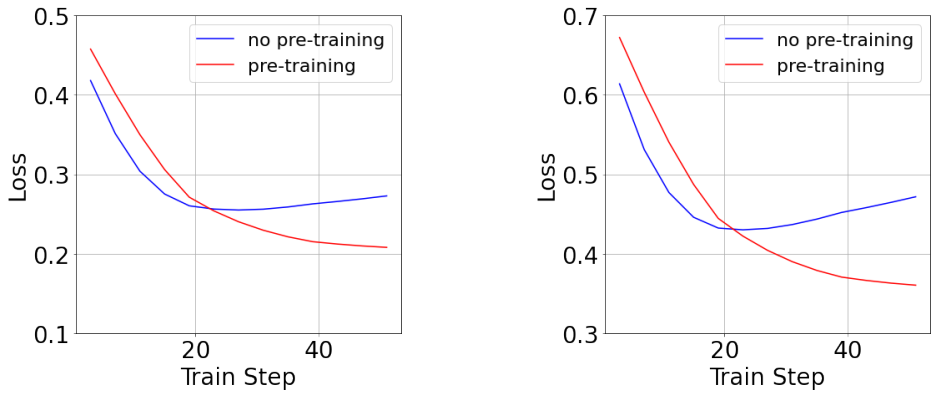
These results indicate that the complete model benefits from the pre-training done to the uni-signal models. It is interesting to analyze how the model gets these benefits. We hypothesize that it could be explained by the two following reasons:

1) The pre-trained uni-signal models are already better. When using single signals, the pre-trained uni-signal models already gave better results than the no pre-trained uni-signal models, as shown in Chapter 3, in particular in Table 3.4. Therefore, as our FCN multi-signal emotion classifier uses the features generated by those uni-signal models, when using the features from the pre-trained models it uses better features than when using the features from no pre-trained models. Using pre-trained uni-signal models that give better features is clearly an advantage that leads to a better performance of the multi-signal modal.

2) The benefits of pre-training are carried over. The benefits of pre-training the uni-signal models are carried over when training the complete architecture. One benefit of pre-training is that the model becomes less prone to overfitting (see Section 3.4.2.3). Therefore, using the pre-trained uni-signal models makes the whole architecture less prone to overfitting.

We further analyze the second reason, as this reason is less evident. It is less evident because the weights of the uni-signal models are frozen when the whole architecture is trained. This means that the uni-signal models cannot become less prone to overfitting. On the other hand, it could be the case that the representations from the uni-signal models carry the benefits of pre-training when they are used as inputs to train the multi-signal emotion classifier (See Figure 4.5). If this is true, the multi-signal classifier that uses representations from pre-trained uni-signal models should be less prone to overfitting. Below, we check this last statement.

Figure 4.6 compares the losses in the validation dataset when training the model using pre-trained uni-signal models (red line) with using uni-signal models that were not pre-trained (blue line). This figure shows the average



(a) Arousal Losses

(b) Valence Losses

Figure 4.6 – Comparison of the losses on the validation set when using pre-trained uni-signal models, in red, compared to not using pre-trained uni-signal models, in blue. We show the losses for arousal (a) and valence (b). The figure shows the average loss across the 10 folds on the AMIGOS dataset.

validation loss across the 10 folds for the AMIGOS dataset. We can see that when using the representations generated by pre-trained uni-signal models to train the multi-signal emotion classifier, the model does not overfit. On the other hand, when not using pre-trained uni-signal models, the model overfits.

These results show that pre-training leads to more robust features, and that in fact, these features carry the benefits of pre-training. Thus, when using them to train a new model, in our case the multi-signal emotion classifier, this new model becomes less prone to overfitting, which is one of the benefits of pre-training.

4.4.2.4 Comparison with the State-of-the-Art

Table 4.4 shows the performance of our model next to other state-of-the-art works that perform emotion recognition using multiple physiological signals. These works use different experimental protocols to evaluate their performance, and therefore they are not directly comparable to each other nor are they comparable to our work. For instance, there is a variety of input segment sizes, input signals, different partitions of data into training and test sets, subject-dependent and independent evaluations, etc. However, all of them recognize categories of arousal and valence, i.e. all of them perform classification of arousal and valence. Therefore, the results in Table 4.4 still give a good idea of the relative performances of the current state-of-the-art,

Model	Signals	Arousal Acc.	Arousal F1	Valence Acc.	Valence F1
		AMIGOS			
GNB [135]	ECG+EEG+EDA	-	0.56	-	0.56
LSTM [120]	ECG+EEG+EDA	0.83	0.72	0.78	0.7
2D-CNN [175]	ECG+EEG+EDA	0.83	0.76	0.84	0.82
2D-CNN [59]	ECG+EDA	0.79	0.75	0.79	0.76
VAE [212]	ECG+EEG+EDA	0.69	0.64	0.67	0.67
VAE [154]	ECG+EDA	0.93	0.95	-	-
Transf. (ours)	ECG+EEG	0.89	0.89	0.85	0.85
		DREAMER			
SVM [102]	ECG+EEG	0.62	0.58	0.62	0.52
GRU [104]	ECG+EEG	0.85	-	0.84	-
DCCA [126]	ECG+EEG	0.89	-	0.91	
Transf. (ours)	ECG+EEG	0.84	0.84	0.80	0.79

Table 4.4 – Comparison of our results with other works. These results are not directly comparable as the experimental protocols are not necessarily the same.

showing that we obtain competitive results.

Table 4.4 illustrates the variety of solutions that have been proposed for the task of emotion recognition from multiple physiological signals. Non-DL methods like GNB and SVM that use engineered features as inputs are presented in Miranda-Correa et al. [135], and Katsigiannis and Ramzan [102]. However, Table 4.4 shows that results tend to be better with DL approaches. Some authors explore the use of recurrent networks, with the usage of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) by Li et al. [120] and Khan et al. [104], respectively. Liu et al. [126] use Deep Canonical Correlation Analysis (DCCA), which is a technique where two FCN are used to obtain representations of two modalities, learning the weights of those FCN by maximizing the correlation between those representations. To take advantage of existing pre-trained models for computer vision tasks, Siddharth et al. [175] and Elalamy et al. [59] use 2D Convolutional Neural Network (2D-CNN) to process images (e.g. spectrograms) generated from the physiological signals. Finally, Yang and Lee [212], and Ross et al. [154] use VAE to generate representations, pre-training the VAEs by encoding and reconstructing the inputs.

Given the variety of evaluation protocols employed by the works pre-

sented in Table 4.4, it is not possible to conclude which approach has the best performance. Nevertheless, as all of those approaches recognize categories of arousal and valence, several conclusions can be drawn from the results shown in that table. First, DL approaches tend to give more accurate results than non-DL approaches, corroborating the ability of DL models to extract and process features from data. Second, some authors have used pre-training techniques, confirming the usefulness of this method when predicting emotions from multiple physiological signals. Third, our pre-trained Transformer-based approach shows results that are on the line with other state-of-the-art works, demonstrating the validity of our approach.

4.5 Conclusions

This chapter presented the second contribution of this thesis, which is a pre-trained Transformer-based technique designed to recognize emotions from multiple physiological signals. Our method is based on pre-training and fine-tuning individual uni-signal models and using late fusion to combine the outputs of those models. This approach addresses Challenge 4.1, since using pre-training produced better results than when no pre-training was used. As discussed in the chapter, the improvement when using pre-training could be explained by the better quality of results of each single-signal model and also because the representations from those models were more robust against over-fitting.

We also tested the usefulness of combining multiple physiological signals and found that in many cases this led to better accuracy and F1-score than using one physiological signal. In particular, there was a performance improvement with the AMIGOS dataset when using ECG and EEG signals at the same time in comparison to using those signals separately. With the DREAMER dataset, the accuracy and F1-score when combining ECG and EEG signals were better than using only EEG signals, and at least as good as using only ECG signals. We hypothesize that these results on the DREAMER dataset can be explained by the low performance of the EEG model when using this dataset.

In this chapter, we used only physiological signals to predict high and low categories of arousal and valence. But in a smart environment there might be other types of signals, like visual and sound signals, that can be combined with those physiological signals to perform emotion recognition. Therefore it is appealing to design an architecture capable of using all those types of signals for the task of emotion recognition. For this reason, we

present in the next chapter an approach capable of using multimodal inputs. This approach also performs time-continuous value-continuous emotion recognition, which we believe will help to produce a better image of the mental well-being of a person.

CHAPTER 5

TIME-CONTINUOUS MULTIMODAL EMOTION RECOGNITION

People express emotions through external manifestations in both verbal and non-verbal manners. Examples of non-verbal communication include facial expressions and speech pitch intensity. In addition, as seen in previous chapters, emotions are also reflected in internal manifestations through different physiological signals. A smart environment may be equipped with cameras, microphones, and other sensors capable of collecting those external and internal manifestations. Therefore, to better monitor the mental well-being of a frail person in such smart environments, it is desirable to automatically infer emotions with the multimodal information coming from such sensors.

This chapter describes a method for recognizing emotions from multimodal inputs, in a time-continuous fashion. Specifically, the chapter presents a model that not only processes physiological signals, like in previous chapters, but uses other modalities, like audio and video, to perform time-continuous emotion recognition.

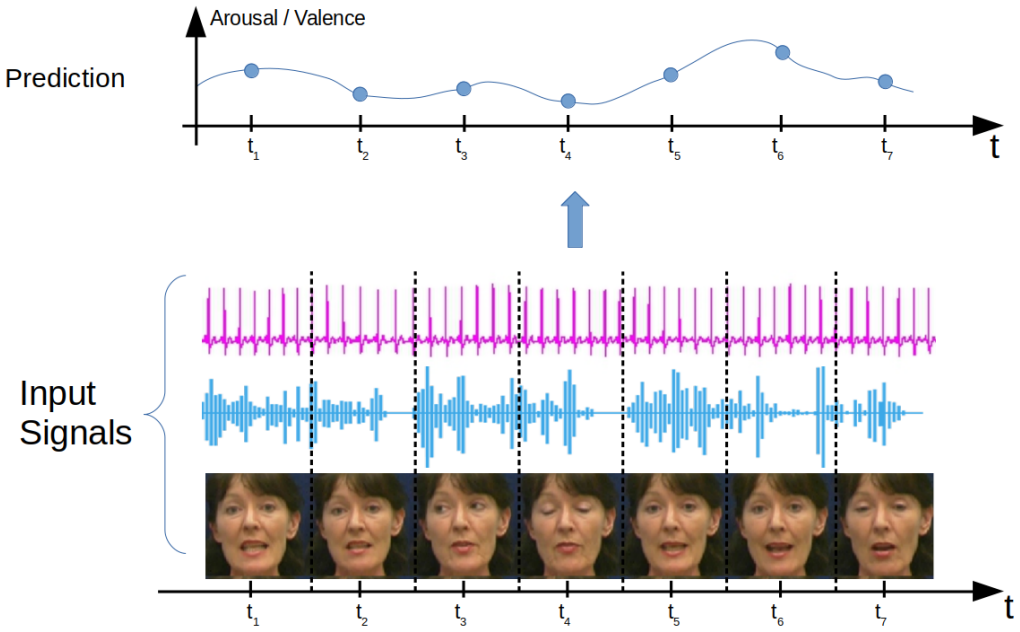


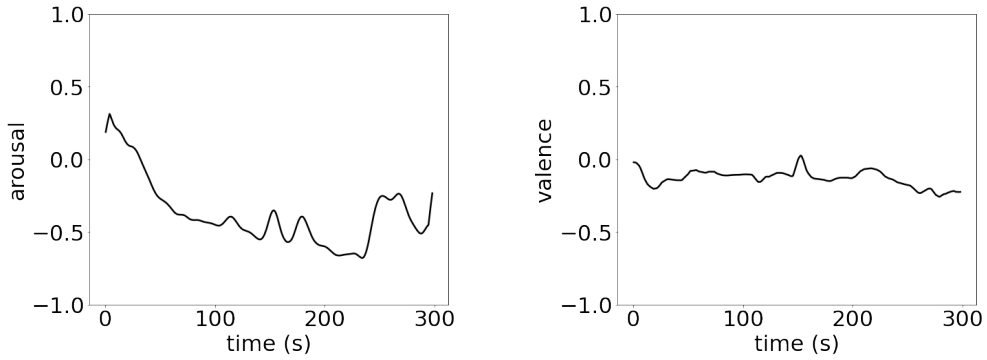
Figure 5.1 – Depiction of the problem addressed in this chapter. From multimodal inputs, we want to predict value-continuous and time-continuous levels of arousal and valence.

This chapter provides a definition of the problem being addressed, as well as the motivations and challenges associated with that problem in Section 5.1, examines the related State-of-the-Art in Section 5.2, introduces our approach for time-continuous emotion recognition in Section 5.3, and shows the experimental results when testing our approach in Section 5.4.

5.1 Problem Definition, Motivations and Challenges

5.1.1 Problem Definition

Our goal is to use the signals from multiple perceptual modalities, including video, audio, and physiological signals, to recognize continuous values (in the [-1, 1] range) of arousal and valence in a time-continuous fashion. Since we want to recognize continuous levels of emotion, this task is a regression problem. Note that throughout this chapter, we refer to this problem as *multimodal continuous emotion recognition*, and we use the word *continuous* to denote value-continuous and time-continuous, and the word



(a) Arousal ground-truth example.

(b) Valence ground-truth example.

Figure 5.2 – Example of time-continuous ground-truth values, sampled at 2Hz.

emotion to indicate either arousal or valence.

Figure 5.1 presents a depiction of the problem addressed in this chapter. The multimodal inputs are temporal sequences of visual, audio, and physiological signals. Specifically, the visual information is a sequence of image frames from a video; the audio is a time-series of audio values, and the physiological signals are also time-series values. Then, emotion recognition is performed at a certain rate, say at each time-step t, such that for every time-step t there is a corresponding image frame (or several image frames), a segment of audio, and a segment of a physiological signal. In our case, we assume that for every time-step, there is a label indicating the arousal and valence values, which means that the inputs and labels are aligned. In addition, we assume that there are no missing modalities, an assumption that will be relaxed in Chapter 6.

In our case, the recognition rate is high enough (2Hz), so there are not large variations of arousal and valence between two consecutive recognized values. For this reason, we talk about time-continuous emotion recognition. Figure 5.2 shows an example of arousal and valence ground-truth values that we aim to recognize.

5.1.2 Motivations

5.1.2.1 Motivation to use Multimodal Inputs

Our motivation to use multimodal inputs comes from the idea that the information in different modalities may be complementary. In fact, in Chapter 4, we showed that using multiple physiological signals can improve the

accuracy of the predicted values from emotion recognition models. In this chapter, we explore this idea further using other modalities besides physiological signals, namely visual, and audio modalities. In our case, audio information captures speech, and visual information captures facial expressions. As detailed in Section 2.2.3, speech and facial expressions are ways that emotion is displayed, thus it is feasible to use them in combination with physiological signals as inputs for an emotion recognition system. Moreover, since these modalities are of different nature and produced by different mechanisms in the body, they should carry different and complementary information, thus helping to improve the accuracy of an emotion recognition model.

Another motivation to use multimodal inputs is that a smart environment may be equipped with sensors to gather these different modalities. Portable sensors could be worn to gather physiological signals, as it was described in previous chapters. In addition to this, cameras and microphones could gather audio and video signals, both of which could be acquired through user interaction with a smart assistant, for example. In this scenario, privacy concerns may arise. We discuss these concerns and other ethical implications of our work in the perspectives provided in Section 7.2.9.

5.1.2.2 Motivation to Perform Continuous Emotion Recognition

Although knowing if an emotion is positive or negative and if its intensity is low or high may be enough to have an idea of the emotional state of a person, we think that to have a better picture of the emotional situation it can be useful to obtain value-continuous levels of arousal and valence. For example, stress can be inferred from the level of arousal and valence, as there is a correlation between them [188].

Moreover, continuously recognizing the emotional state of a person can be a way to improve the monitoring of his or her emotional well-being. This results from the fact that predicting emotions in a time-continuous manner can show the emotional variations that a person feels across time, which can give better insights into the emotional state of this person. For example, if a decision has to be made using the inferred emotional state of a frail person, it would be better if this decision is made following several consecutive predictions, rather than basing this decision on a single prediction.

5.1.3 Challenges

Three main challenges have to be addressed to construct a system for multimodal continuous emotion recognition. The first challenge is how to model the temporal dependencies inside the input sequence of each modality. The second challenge is how to fuse the information from the different modalities. The third challenge is how to take past predictions into account when performing the current prediction.

Since each modality is a temporal sequence, there are temporal dependencies inside each modality that are important to identify and model in order to efficiently use the information from that modality. With this, we define the first challenge of this chapter as follows:

Challenge 5.1. *How to model the temporal dependencies present in each modality.*

Using multimodal inputs is advantageous since the information from the different modalities may be complementary, leading to better results from the model. However, it is important that the model is able to extract and take advantage of this complementarity, also taking into account that part of the information may be redundant. Thus, the second challenge addressed in this chapter is:

Challenge 5.2. *Designing a model capable of aggregating multimodal information, taking advantage of complementary information, and discarding redundant information.*

Since emotions will be predicted in a time-continuous manner, it is important to use past predictions to infer the current emotional state, because this past information may help to obtain better results. Changes in emotion are not instantaneous, thus there is a relation between the current and past emotional states. With this, we define the third challenge as follows:

Challenge 5.3. *How to use past emotion predictions when inferring the current emotion.*

5.2 Existing Techniques for Multimodal Continuous Emotion Recognition

As stated above, when doing multimodal continuous emotion recognition, there are three main challenges that need to be addressed. The first

one is how to model the temporal dependencies from the input sequence (Challenge 5.1), the second one is how to fuse the different modalities (Challenge 5.2), and the third is how to take past predictions into account (Challenge 5.3). Regarding the first challenge, in the literature there are two main methods: recurrent models [93, 94, 193] and attention-based models [38, 87, 219]. For the second challenge, some approaches use early and late fusion [128, 220], while other authors have studied the use of attention mechanism for multimodal fusion [143, 190, 219]. We did not find in our literature review contributions that directly address the challenge of taking previous predictions into account, although using recurrence does this in a way.

5.2.1 Modelling Time-Continuous Information for Emotion Recognition

When recognizing time-continuous values of emotion from multimodal inputs, the input data are sequential, containing information across time. Therefore, it is important to model the underlying temporal dynamics present in these data. One way of doing this is to use architectures designed to process sequences. In the literature, this has been done mainly using two architectures: Recurrent Neural Network (RNN) [93, 94, 193], and attention mechanisms [38, 87, 219]. The following subsections discuss in more detail some papers that present these approaches.

5.2.1.1 Recurrent Approach for Continuous Emotion Recognition

In [93], Huang et al. present an approach for continuous emotion recognition with RNNs, using audio, visual, and text modalities. To process the input sequences, they first perform a data augmentation technique by segmenting the original sequence into smaller overlapping segments. This way, they obtain a larger quantity of training samples. In addition, since the samples are smaller than the original sequences, they are more suitable to be processed to model the temporal dynamics inside those sequences. The dataset they use contains sound recordings of the target speaker and an interlocutor. For this reason, Huang et al. [93] append to the sound features a marker to differentiate from who the features come from. This marker is a 1 for the main speaker and a 0 for the interlocutor.

To process the signals and model the temporal relations inside them, Huang et al. [93] use a Long Short-Term Memory (LSTM) network. Before feeding the modality features into the LSTM, they are processed with average

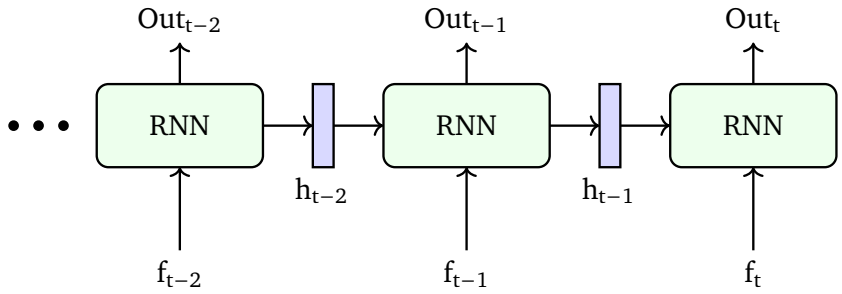


Figure 5.3 – Recurrent Neural Network (RNN). Note that when obtaining the output Out$_t$, all the past information is contained in the hidden state vector h_{t-1}.

pooling in the temporal dimension to achieve an additional short level of temporal modeling. A time delay is added, shifting the features with respect to the labels, to overcome the annotation delay present in the dataset they use.

To fuse the information from the different modalities, the authors consider feature-level fusion (i.e. early-fusion) and decision-level fusion (i.e. late-fusion). Feature-level fusion is performed by concatenating the features before they are fed into the LSTM network, while decision-level fusion is done by training an individual model for each input modality, and then the outputs of the different models are concatenated and processed with support vector regression to infer the final emotion value. Feature-level fusion and decision-level fusion obtain comparable results when predicting arousal, while feature-level fusion achieves better results for valence prediction.

Although this work shows that using recurrent models, like LSTM networks, are useful for time-continuous emotion recognition, RNNs convey the information of past interactions into a single hidden state vector, as depicted in Figure 5.3. Therefore, long-range interactions are difficult to model with an RNN, even with the improvements that LSTM networks bring by controlling the information flow. One solution to this is the use of attention-based approaches that allow the direct incorporation of past (or future) information. One of the most successful attention-based models is the Transformer, which was discussed in detail in Section 2.3.

5.2.1.2 Attention-Based Approach for Continuous Emotion Recognition

Chen et al. [38] introduce a model for multimodal continuous emotion recognition. They propose to use attention layers to model the tempo-

ral dynamics within each modality. In addition, they also use attention mechanisms to model the inter-modality interactions.

The proposed approach first extracts the features for each modality, then processes these features with a 1D Convolutional Neural Network (1D-CNN) to aggregate local context information. The output features from the 1D-CNN are processed by a stack of Multimodal Attention Modules and Temporal Attention Modules, and the output of these modules is processed with a Fully-Connected Network (FCN) to predict the emotion values.

The Multimodal Attention Module in the approach of Chen et al. [38] is designed to model the inter-modality interactions, while the Temporal Attention Module is in charge of modeling the temporal intra-modality interactions. In general terms, the Temporal Attention Module works by having as input, for each modality, a sequence of time-steps, thus each time-step attends the other time-steps within the same modality. On the other hand, the Multimodal Attention Module works by having as input, for each time-step, a sequence of modality features, thus each modality attends the other modalities within the same time-step.

Regarding the performance of this approach, Chen et al. show that, in several instances, their results are better than other solutions that use recurrence. In the cases where other approaches are better, their results are not far away, remaining competitive. In summary, the work of Chen et al. shows that it is possible to use attention-based models for the task of continuous emotion recognition. However, they do not use past predictions to infer the current emotional state.

In an attempt to further improve the performance of continuous emotion recognition systems, some authors have combined attention modules with RNNs. The following subsection describes in detail a contribution that uses this type of approach.

5.2.1.3 Hybrid Models for Continuous Emotion Recognition

In their work, Wu et al. [208] combine a Transformer encoder with LSTM networks to predict continuous values of valence from video, audio and text. The first step of their approach is to process the features extracted from the different modalities with 1D-CNNs. The fusion of modalities is done by concatenating the features of each modality obtained from the 1D-CNNs. The multimodal sequence is then fed into a Transformer encoder, which produces an intermediate sequence. This intermediate sequence is

further processed by an LSTM, which produces an output sequence that is fed into a FCN to obtain the predicted values.

In their paper, Wu et al. [208] demonstrate that the combination of attention and recurrent mechanisms gives more accurate results than using those mechanisms individually. Moreover, sometimes the performance that their approach obtains is close to a human-level benchmark. These results show the utility of using attention mechanisms to model the temporal information in the inputs. In addition, the results show that using a vanilla implementation of an attention model, like the Transformer encoder, might not be enough to effectively process the inputs in the task of emotion recognition, and architecture adaptations, like the addition of the LSTM network, might be necessary.

5.2.2 Combination of Modalities for Multimodal Emotion Recognition

Having reviewed some approaches to model the temporal dynamics in the sequential inputs used for multimodal continuous emotion recognition, we now focus on another important aspect of multimodal processing: how to combine the information from different modalities. The different approaches found in the literature to combine modalities can be divided into two classes. The first class consists in using early or late fusion, which is typically done by concatenating the input features of each modality before they are fed into the model (early fusion) or concatenating the outputs of individual-modality models (late fusion). A description of these two types of approach can be found in Section 4.2.1. The second class consists of approaches where the combination of modalities goes beyond simple concatenation, but different components within the model itself are in charge of aggregating the information. We call this type of approach model fusion. The following subsections describe some contributions that present these fusion approaches.

5.2.2.1 Early and Late Fusion for Multimodal Emotion Recognition

In [220], Zhang et al. make a comparison between early and late fusion for the multimodal continuous emotion recognition task. Specifically, they work with audio, visual, text and physiological inputs, extracting different types of features from each of them and trying several combinations of these features. Their model consists of an LSTM network that, in the case of early fusion, is fed with the concatenation of the different input features. In the

case of late fusion, they concatenate the outputs of individual LSTM models and process the concatenated sequence with a second-level LSTM.

Zhang et al. [220] did several experiments using different combinations of modalities/features, testing early and late fusion, always using the same hyper-parameters for their architecture. When comparing the results between early and late fusion, they could not find an approach that gives more accurate results than the other in all the cases. On the contrary, the performance depends on the number and type of modalities, as well as the type of features that are used from each modality. For example, the authors found that when audio and visual modalities are used, early fusion gives more accurate results. On the other hand, when audio and text are used, the late fusion approach gives more accurate results. In the case of audio and physiological features, early or late fusion gives more accurate results depending on what type of features of each modality are used.

Additionally to testing different fusion methods, Zhang et al. [220] compare using a multimodal approach with using a single modality. The authors experimentally show that using a multimodal approach can lead to performance improvements, demonstrating that in some cases the information in the different modalities is complementary.

Several other authors [73, 139, 190] have experimented with more sophisticated ways to aggregate multimodal information beyond early and late fusion. Below, some of those contributions are reviewed.

5.2.2.2 Multimodal Transformer

The idea behind the multimodal Transformer is to concatenate the sequence of different modalities one after the other and then feed this multimodal sequence into a Transformer encoder, as indicated in Figure 5.4. This way, the attention mechanism of the Transformer can model the intra-modality dependencies of each modality, and at the same time, they can model the inter-modality dependencies between modalities. One contribution that uses this approach is the work of Gabeur et al. [73]. Although the aim of this work is not emotion recognition, it is still interesting to analyze as an example of a multimodal Transformer. Gabeur et al.'s approach processes multimodal signals by first extracting different types of features of each modality using a pre-trained model. They call each type of feature an *expert*. To obtain an aggregated representation of each expert, they use a vector F^m_{agg} that is appended to the beginning of the feature sequence of each expert m. This is similar to the CLS token used in the architecture

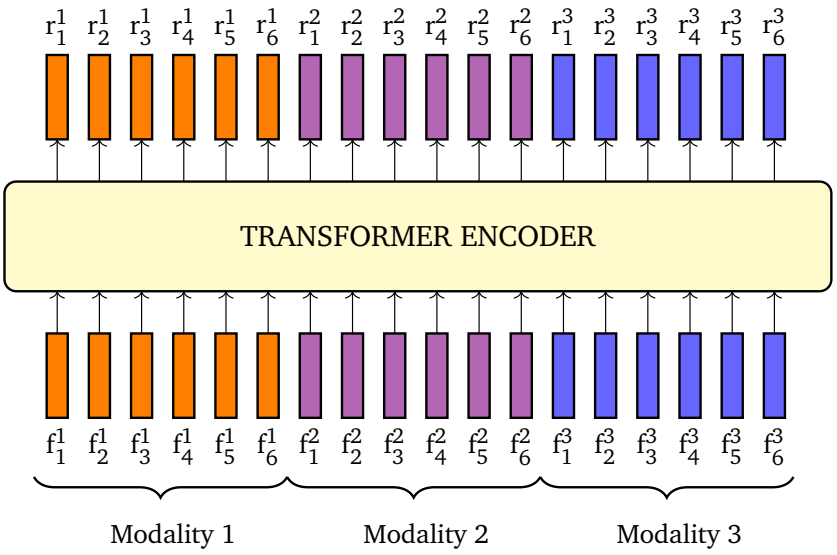


Figure 5.4 – Using a Transformer Encoder with multimodal inputs. The different modalities are concatenated, and the resulting sequence is fed into a Transformer encoder.

presented in Chapter 3. Denoting each expert feature as F_t^m, the sequence of input features has the form

$$F = [F_{agg}^1, F_1^1, \ldots, F_T^1, \ldots, F_{agg}^M, F_1^M, \ldots, F_T^M], \tag{5.1}$$

where M is the number of experts, and T is the sequence length of every expert. To allow the model to distinguish between experts, M embeddings are learned, one for each expert. These embeddings are added to the features of the corresponding expert. In addition, as is typically done in Transformer-based architectures, temporal embeddings are added so the model is provided with information about the order of the sequence. The final sequence is fed into a Transformer encoder, and then the authors use the outputs corresponding to the aggregated feature F_{agg}^m as representations of each expert.

This work shows that attention-based models are capable of generating representations of each modality that aggregate the intra-modality temporal information and the inter-modality information. Nevertheless, Nagrani et al. [139] argue that while it may seem better to have a free flow of information across modalities, it is not necessary to have this flow of information in all the layers of the model because part of the information might be redundant. For this reason, Nagrani et al. propose to restrict this flow by using a small set of fusion units through which the interchange of information between modalities must pass.

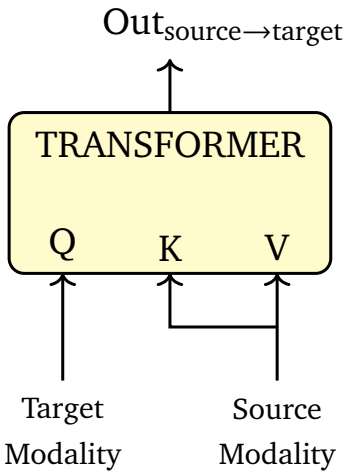


Figure 5.5 – A Crossmodal Transformer [190] incorporates information from the source modality into the target modality by taking the key and value vectors from the source and the query vectors from the target.

From the works of Gabeur et al. [73] and Nagrani et al. [139], we can conclude that the multimodal Transformer is capable of modeling interactions between modalities. Specifically, both of these architectures use self-attention because each position in the input sequence attends to the input sequence itself (see Section 2.3.1.3). However, there is a different approach that might be used, which is using cross-attention instead of self-attention. The following subsection details this approach.

5.2.2.3 Cross-Attention for Multimodal Fusion

The idea behind using cross-attention for multimodal fusion is to use multiple pairwise crossmodal Transformers, with each of these Transformers reinforcing a target modality with information from a source modality. This idea was introduced by Tsai et al. in [190]. As described in section 2.3, the attention mechanisms inside a Transformer require query, key, and value vectors (Q, K, and V, respectively). Then, the idea of Tsai et al. [190] is to take the query from one modality, the target modality, and the key and value vectors from another modality, the source modality, as depicted in Figure 5.5. The advantage of this approach is that the source and target modalities do not need to be aligned. To illustrate what we mean by alignment, we can take, for example, a video and a sentence that corresponds to what a person says in that video. The first word may correspond to the first seven frames, the second to the next five frames, the third to the next ten frames, etc.

Thus, in this example, aligning the video and the language inputs means identifying what frames correspond to what words.

Regarding the usage of cross-attention for multimodal fusion for continuous emotion recognition, we detail the contribution of Huang et al. [95], where they use as inputs audio-visual information. In their work, they first process audio and visual information individually using a Transformer encoder, and then, to combine the information from those two modalities, they use a crossmodal Transformer. Huang et al. [95] choose only to incorporate information from audio to the visual modality and not do it also the other way around. This means that in their crossmodal Transformer, the query vector comes from the visual modality, while the key and value vectors come from the audio modality. The final step in their approach is to use a linear layer to process the outputs of the crossmodal Transformer and obtain the predicted values of emotion.

Using a crossmodal Transformer is an interesting approach that can be especially useful when the different modalities are not aligned. However, the problem is that this approach allows only a pair-wise exchange of information, thus if a complete information flow between several modalities is required, the number of crossmodal Transformers needed is $2\binom{M}{2}$, where M is the number of modalities. Nevertheless, the contributions reviewed in this subsection further confirm the value and utility of using attention to aggregate information from different modalities.

5.2.3 Discussion

In order to predict continuous values of emotion, it is necessary to process sequential inputs. In the literature, we have found two main trends to process sequences: the first one is to use RNN, and the second one, which is more recent, is to use attention. Attention-based approaches have the advantage that they are better adapted to model the long-range relations that may be present in the inputs. In addition, these models can process the inputs in parallel because to obtain the output at a specific time-step it is not necessary to have obtained the output from the previous step. Moreover, the representations generated with attention-based models incorporate weighted information from the whole input, meaning that it identifies the important parts of the input signal, giving more weight to them when building the representation. Regarding the fusion of information from multiple modalities, using attention has the advantage that such architectures are capable of aggregating information from the different modalities, but they

do in a way that more weight is given to the most important ones.

For these reasons, in our architecture, we use an attention-based approach to model the temporal dynamics of the inputs and to aggregate the information from the different modalities. Since we are not concerned with unaligned modalities, we do not use a crossmodal Transformer. Instead, we use a novel approach that aggregates the information from the different modalities using the cross-attention from a Transformer decoder, incorporating at the same time past predictions in an autoregressive manner. With our approach, we take advantage of the characteristics of an attention-based model while also restricting the information between different modalities, which, as suggested by Nagrani et al. [139], should increase the performance of the model.

5.3 Multimodal Transformer for Emotion Recognition

This section presents our approach to addressing the problem of multimodal continuous emotion recognition. In our approach, we use the same architecture to recognize both arousal and valence, although we train one independent model to recognize arousal and another to recognize valence. For this, in this section, we often use the terms to recognize (or predict) an emotion to refer to recognize (or predict) values of either arousal or valence.

Figure 5.6 shows our architecture, which follows an encoder-decoder approach, similar to the one presented in the original Transformer paper [196]. The encoder, which we call Multimodal Transformer Encoder (MMTE), takes elements from current literature and is capable of producing representations using all the modalities as inputs. The multimodal representations generated by the encoder are processed and aggregated by the decoder, called Autoregressive Multimodal Transformer Decoder (AMMTD). The AMMTD produces the representation d_t that is used to predict the emotion value at time-step t with the help of a FCN network called Emotion Regression Network (ERN). We design this decoder such that it aggregates information from the multimodal representations given by the encoder, in a way that more weight is given to the most important modalities. To generate the representation d_t, the decoder also considers the generated representations from previous time-steps. The following subsection explains in detail the two main components of our architecture: the MMTE and the AMMTD.

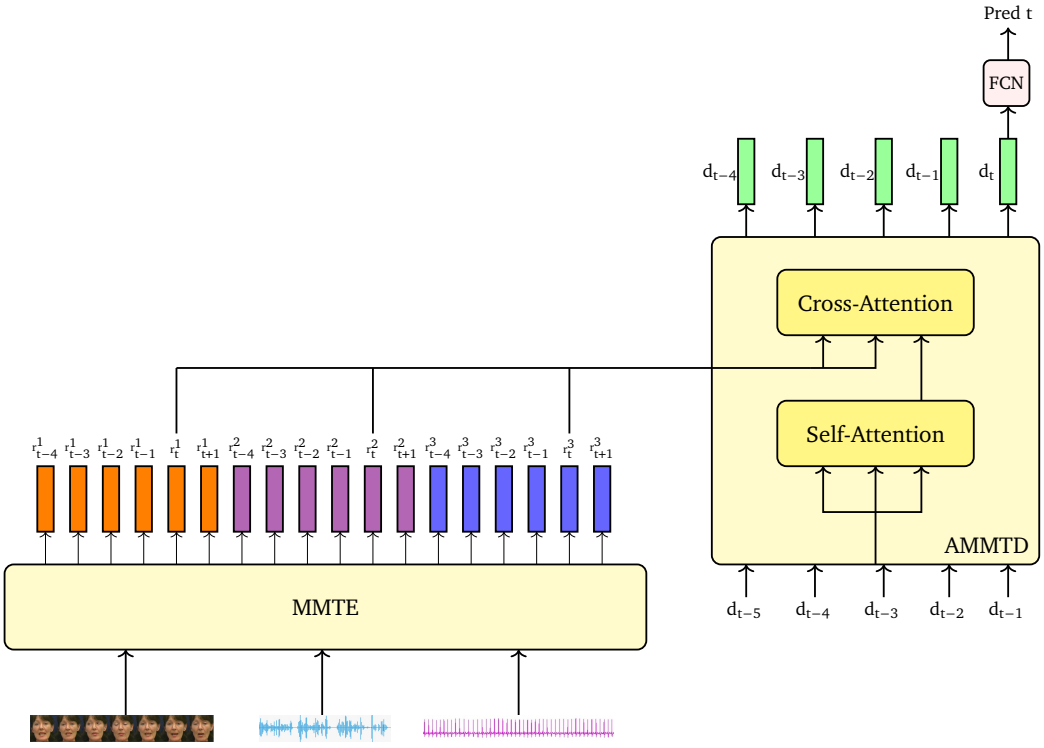


Figure 5.6 – General architecture of our approach for multimodal emotion recognition.

5.3.1 Multimodal Transformer Encoder (MMTE)

To design the MMTE, which is depicted in Figure 5.7, we took some inspiration from the work of Gabeur et al. [73]. The encoder takes as inputs features extracted from the different modalities. We discuss the features used in Section 5.4.1.3. The MMTE models the temporal dynamics inside each modality, addressing the Challenge 5.1.

The first step in the MMTE architecture is to process each modality individually using a Temporal Convolutional Network (TCN) [15] to model local temporal information, similarly to what is done by Chen et al. [38]. TCNs are Convolutional Neural Networks (CNNs) adapted for sequence modeling, and we use them because the input is a sequence of features. Our architecture uses independent TCNs for each modality. If we define the feature corresponding to modality m at time-step t as $x_t^m \in \mathbb{R}^{d_{\text{modality}}}$, then the input sequence for modality m will be $[x_1^m, \ldots, x_T^m]$, where T is the length of the sequence. That being said, when using the TCN to process the input corresponding to modality m, we have:

$$[a_1^m, \ldots, a_T^m] = \text{TCN}^m([x_1^m, \ldots, x_T^m]), \tag{5.2}$$

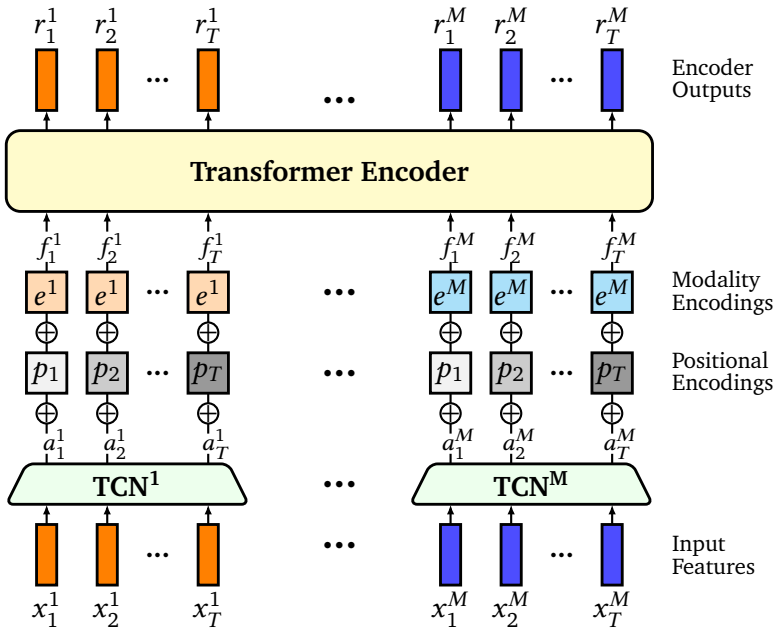


Figure 5.7 – Multimodal Transformer Encoder (MMTE).

where $a_t^m \in \mathbb{R}^{d_{\text{model}}}$. Note that for all modalities, the TCN outputs will have a common size d_{model}.

The outputs given by the TCN will be processed by the Transformer. To do this, it is necessary to add positional encodings to allow the Transformer to take into account the actual order of the sequence [196]. The positional encodings that we use are vectors that are learned during training. Specifically, the model learns a vector $p_t \in \mathbb{R}^{d_{\text{model}}}$ for each input position, giving the sequence $[p_1, \ldots, p_T]$. Then, the sequence obtained is:

$$[a_1^m + p_1, \ldots, a_T^m + p_T]. \tag{5.3}$$

We use learned positional encodings because we did preliminary experiments that showed that in our case they worked better than fixed positional encodings. See Section 2.3.2.3 for more information about positional encodings.

To process cross-modality information, the Transformer must discern between each modality. To achieve this, we adopt the approach of Gabeur et al. [73] and use modality encodings. Like positional encodings, these encodings are vectors learned during training. Specifically, for each modality m, an encoding $e^m \in \mathbb{R}^{d_{\text{model}}}$ is added to the input. The output after doing

this is

$$[f_1^m, \ldots, f_T^m] = [a_1^m + p_1 + e^m, \ldots, a_T^m + p_T + e^m] \tag{5.4}$$

Note that f_t^m is the vector that corresponds to modality m at time-step t, with $f_t^m = a_t^m + p_t, +e^m$.

We concatenate the sequences from all modalities to form a single sequence. If we have M modalities, the concatenated input sequence is

$$[f_1^1, \ldots, f_T^1, \ldots, f_1^M, \ldots, f_T^M]. \tag{5.5}$$

The concatenated sequence is processed using a Transformer encoder, which produces representations r_t^m given by

$$\begin{gathered}[r_1^1, \ldots, r_T^1, \ldots, r_1^M, \ldots, r_T^M] = \\ \text{Transformer Encoder}([f_1^1, \ldots, f_T^1, \ldots, f_1^M, \ldots, f_T^M]).\end{gathered} \tag{5.6}$$

We employ a bidirectional attention mask at the input of the Transformer encoder, similar to what is done in the work of Chen et al. [38]. When the Transformer encoder is processing an input f_t^m, this mask *hides* the inputs that are farther than *mask_length* positions in the future and in the past. This means that to produce the representation r_t^m, the Transformer attends the sequence $[f_{t-mask\_length}^m \ldots f_{t+mask\_length}^m]$. This allows the model to concentrate on recent information and not to worry about information too far in time that probably does not influence the current emotional state.

5.3.2 Autoregressive Multimodal Transformer Decoder

The decoder in our architecture uses the representations generated by the encoder to predict the values of arousal and valence. This decoder needs to address Challenge 5.2 and Challenge 5.3. This means it needs to aggregate the representations of the different modalities while restricting the information flow to deal with redundant information, and it has to take previous predictions into account to determine the current emotion. To do this, we design the Autoregressive Multimodal Transformer Decoder (AMMTD), which is described below.

To understand how the AMMTD is built, we give a brief description of the Transformer decoder [196], which is explained in more detail in Section 2.3. The Transformer decoder is constructed of stacked Transformer Decoder Layers (TDLs). Each TDL is composed of a Multi-Head Self-Attention (MHSA) module, followed by a Multi-Head Cross-Attention (MHCA) module, and

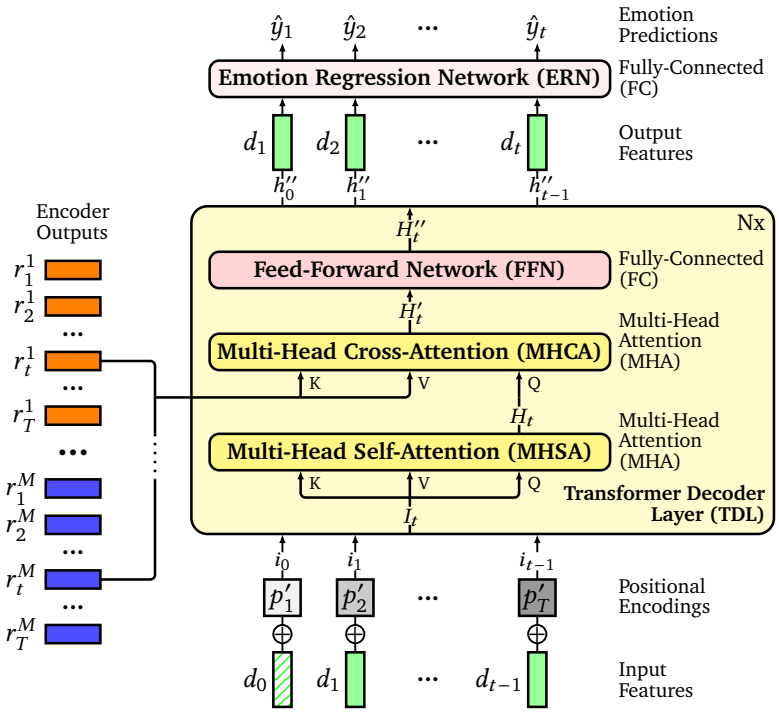


Figure 5.8 – Autoregressive Multimodal Transformer Decoder (AMMTD). State of the decoder when predicting the emotion value at time t.

followed by a fully-connected Feed-Forward Network (FFN). Residual connections are used around each of these three components. The MHSA and MHCA implement a Multi-Head Attention (MHA) mechanism that projects a query vector q from a given position to a key vector k from another position to determine the attention (i.e. the weight) given to a value v associated with the position of k. The final value is the weighted sum of the v vectors from the different positions. We denote the MHA mechanism as

$$\text{MHA}(Q, K, V), \tag{5.7}$$

where the three parameters Q, K, and V indicate the sequence used as query, key, and value, respectively.

The AMMTD, shown in figure 5.8, is composed of a stack of TDLs, followed by a ERN. We use autoregression to address Challenge 5.3, taking into account past predictions. This means that the previously generated outputs are used as inputs to the AMMTD. Note that we do not use the outputs of the ERN, i.e. the predicted emotion values $\hat{y}$, since these are scalar values and the inputs of the AMMTD module must be vectors. For this reason, we use the features generated by the top TDL. Specifically, at

the moment of predicting the emotion value at time-step t, the sequence $[d_1, \ldots, d_{t-1}]$ should already have been generated by the stack of TDLs. Then, the AMMTD input is

$$[d_0, d_1, \ldots, d_{t-1}], \tag{5.8}$$

where $d_0 \in \mathbb{R}^{d_{\text{model}}}$ is a randomly initialized vector.

In the same way as it was done for the encoder, the decoder uses positional encodings $p'_t \in \mathbb{R}^{d_{\text{model}}}$ that are added to the input before it is fed to the stack of TDLs. These positional encodings are learned during training. With this, when performing the prediction at time-step t, the input sequence $I_t = [i_0, i_1, \ldots, i_{t-1}]$ with $I_t \in \mathbb{R}^{t \times d_{\text{model}}}$ becomes

$$I_t = [d_0 + p'_0, d_1 + p'_1, \ldots, d_{t-1} + p'_{t-1}]. \tag{5.9}$$

The sequence I_t from Expression 5.9 is processed by the stack of TDL. As described in previous paragraphs, the three components of a TDL are the MHSA, the MHCA, and the FFN. We now describe how we adapt these components in our architecture, particularly to address the Challenges 5.2 and 5.3.

5.3.2.1 Multi-Head Self-Attention (MHSA)

Inside the TDL, the features are first processed by the MHSA module. This module uses self-attention, meaning that it integrates information from its own input. In other words, the query, key, and value used for the MHA layers inside the MHSA come from the input sequence. Using Expression 5.7, the output of the MHSA module is

$$H_t = [h_0, h_1, \ldots, h_{t-1}] = \text{MHA}(I_t, I_t, I_t). \tag{5.10}$$

Since the sequence I_t is built with past outputs of the decoder (see Expressions 5.8 and 5.9), it means that for the model to generate the sequence H_t, the MHSA module attends the past generated outputs. This is the way past outputs are taken into account, thus addressing the Challenge 5.3.

5.3.2.2 Multi-Head Cross-Attention (MHCA)

The MHCA module is used to incorporate the information from the multimodal signals. For this, the MHCA module has as input the sequence H_t and

attends to the encoder outputs. In other words, the MHA inside the MHCA uses as query the sequence H_t, and as key and value the representations generated by the MMTE module. This means the attention mechanisms in the MHCA are used to aggregate the information from the different modalities. To be more precise, when predicting the emotion value at time-step t, the MHCA attends only to the representations of each modality corresponding to this time-step. Thus, the output sequence of the MHCA is

$$H'_t = \text{MHA}([h_0, h_1, \ldots, h_{t-1}], [r_t^1, \ldots, r_t^M], [r_t^1, \ldots, r_t^M]), \tag{5.11}$$

where the sequence $[r_t^1, \ldots, r_t^M]$ is obtained using Expression 5.6 and then extracting the representations of each modality at time t. This is the way we use the attention mechanisms of the MHCA to aggregate the information from the different modalities, thus addressing Challenge 5.2.

Note that when predicting the emotion value at time t, instead of making the model attend the encoder outputs for that same time-step ($[r_t^1, \ldots, r_t^M]$), we could have instead made the model attend all encoder outputs ($[r_1^1, \ldots, r_T^1, \ldots, r_1^M, \ldots, r_T^M]$). However, as it will be shown in the Experiments section, we found it more effective to have the MHCA module focus only on finding the best weighting between the different modalities, avoiding the MHCA having to weigh which other time-steps in the different modalities might be important. Moreover, having the MHCA module focus on the current time-step restricts the information flow between modalities, forcing the shared representation to condense the most significant information, a technique that has been demonstrated to be beneficial by Nagrani et al. [139]. This flow restriction is also important to address Challenge 5.2.

5.3.2.3 Feed-Forward Network (FFN)

The last component of the TDL processes each element of the sequence H'_t through a fully-connected FFN, applied independently to each position, thus we have:

$$H''_t = \text{FFN}(H'_t). \tag{5.12}$$

The sequence H''_t is the input of the next layer in the TDL stack. Concretely, the next TDL in the stack uses as input $I_t = H''_t$ and implements Expressions 5.10, 5.11, and 5.12. If the sequence H''_t is generated by the top TDL, H''_t becomes the newly generated sequence $[d_1, \ldots, d_t]$ that will be used as input for the AMMTD to predict the emotion value for the next time-step $t + 1$,

After the model has generated the complete output sequence $D = [d_1, \ldots, d_T]$, the final step is to process this sequence with the ERN. As depicted in Figure 5.8, the ERN consists of a FCN that processes independently each element of the sequence D, predicting the emotion values for each time-step and generating the resulting sequence $[\hat{y}_1, \ldots, \hat{y}_T]$.

5.3.3 Expected Results

5.3.3.1 Results Expected from Attention Mechanisms

Our approach for processing multimodal inputs is to use the cross-attention mechanism of a Transformer decoder to weigh the importance of each modality. But there are other alternatives, like concatenating the representations from the different modalities generated by the MMTE to obtain a fused feature, and then processing this fused feature to obtain the predicted emotion. In that case, the importance given to each modality is fixed once the architecture has been trained. On the other hand, using attention can dynamically weigh the importance of each modality according to the input. With this in mind, the following result is expected:

Expected Result 5.1. *Using attention mechanisms to aggregate multimodal information should produce more accurate results when recognizing continuous values of arousal or valence than fusing the multimodal information without dynamically weighing the importance of each modality.*

5.3.3.2 Expected Results on Cross-Attention Length

If the MHCA module attends several time-steps of each modality, this module not only weighs the importance of each modality, but it attends to the temporal information inside each modality. Thus, how many time-steps of each modality are attended should impact the performance of the model. In other words, we expect the following:

Expected Result 5.2. *Changing how many time-steps the MHCA module attends should impact the performance of the model, having an optimal value between attending only 1 time-step (the module concentrates only on weighing the importance of each modality), and attending all the time-steps (the module evaluates the importance of each modality and models temporal dependencies).*

5.4 Experiments

This section shows the results of testing our approach for multimodal continuous emotion recognition, describing the experimental setup in Section 5.4.1, and showing the obtained results in Section 5.4.2. As mentioned before, in our experiments we use the same architecture to predict arousal and valence, although we train one model for arousal and another for valence.

5.4.1 Experimental Setup

5.4.1.1 Metrics

Different from previous chapters where we addressed a classification problem, the problem addressed in this chapter, continuous emotion recognition, is a regression problem. Therefore, we use two metrics adapted for this type of problem: Concordance Correlation Coefficient (CCC) and Root Mean Square Error (RMSE), which are detailed below.

Concordance Correlation Coefficient

The CCC metric, introduced by Lawrence Lin in [123], is a measure of the agreement between the ground-truth values and the predicted values. Given T pairs of predictions and ground-truth values $(\hat{Y}, Y) = (\hat{y}_t, y_t)_{t=1}^{T}$, the CCC measures the correlation between values $\hat{Y}$ and Y, taking into account how close these pairs are to the line $\hat{y}_t = y_t$ (perfect prediction). The metric is built like this because $\hat{Y}$ and Y can have a perfect correlation even in cases when $(\hat{y}_t \neq y_t)_{t=1}^{T}$. The idea behind the CCC is to calculate the expected value of the squared difference between $\hat{Y}$ and Y, which is also two times the expected squared perpendicular distance D of each pair $(\hat{y}_t, y_t)$ to the line $\hat{y}_t = y_t$. With this, in his paper, Lin proposed:

$$\begin{aligned} E[(\hat{Y} - Y)^2] = E[2D] &= (\mu_{\hat{Y}} - \mu_Y)^2 + \sigma_{\hat{Y}}^2 + \sigma_Y^2 - 2\sigma_{\hat{Y}Y} \\ &= (\mu_{\hat{Y}} - \mu_Y)^2 + \sigma_{\hat{Y}}^2 + \sigma_Y^2 - 2\rho\sigma_{\hat{Y}}\sigma_Y, \end{aligned} \tag{5.13}$$

where μ, σ^2, and σ represent the mean, variance, and standard deviation, respectively, for $\hat{Y}$ and Y as indicated by the subindices; $\sigma_{\hat{Y}Y}$ is the covariance between $\hat{Y}$ and Y, and ρ is the Pearson correlation coefficient.

In the case that all the predictions are perfect, $E[(\hat{Y} - Y)^2]$ would be 0. To scale the CCC metric so it fits in the range of $[-1, 1]$, Lin proposes the

following normalization:

$$\begin{aligned} \text{CCC} &= 1 - \frac{E[(\hat{Y} - Y)^2]}{E[(\hat{Y} - Y)^2 | \rho = 0 \text{ (No correlation)}]} \\ &= 1 - \frac{(\mu_{\hat{Y}} - \mu_Y)^2 + \sigma_{\hat{Y}}^2 + \sigma_Y^2 - 2\rho\sigma_{\hat{Y}}\sigma_Y}{(\mu_{\hat{Y}} - \mu_Y)^2 + \sigma_{\hat{Y}}^2 + \sigma_Y^2} \\ &= \frac{2\rho\sigma_{\hat{Y}}\sigma_Y}{\sigma_{\hat{Y}}^2 + \sigma_Y^2 + (\mu_{\hat{Y}} - \mu_Y)^2}. \end{aligned} \tag{5.14}$$

With this formulation, a CCC value of 1 indicates perfect agreement, a value of -1 indicates perfect reversed agreement, and a value of 0 indicates that there is no agreement (i.e. no correlation, having $\rho = 0$).

To calculate the CCC metric for T pairs $(\hat{y}_t, y_t)$, the sample counterparts of the terms of Expression 5.14 can be used:

$$\text{CCC} = \frac{2S_{\hat{Y}Y}}{S_{\hat{Y}}^2 + S_Y^2 + (\bar{\hat{Y}} - \bar{Y})^2}, \tag{5.15}$$

where $S_{\hat{Y}Y}$ is the sample covariance between $\hat{Y}$ and Y, $\bar{\hat{Y}}$ and $\bar{Y}$ represents the sample mean, and $S_{\hat{Y}}^2$ and S_Y^2 are the sample variances of $\hat{Y}$ and Y respectively.

In our case, when evaluating our approach, we calculate the CCC for each sample using Expression 5.15, and the final result is the average of the obtained CCC values.

Root Mean Square Error

The CCC metric gives an idea of how well the predicted values correlate to the real values. To complement this metric, the RMSE is also used. This way, it is possible to present a measurement of the difference between the predicted and ground-truth values. Specifically, we compute the RMSE for each sample, and then calculate the average of those results.

5.4.1.2 Evaluation Dataset

To evaluate our model, we use the Ulm-Trier Social Stress Test (ULM-TSST) dataset, which was presented for the Muse 2021 Challenge [179, 180] and was also used in the Muse 2022 Challenge [42]. This dataset includes video, speech, text, and physiological data collected from volunteers during

a stressful situation emulating a job interview, following the Trier Social Stress Test (TSST) protocol [107]. Data are collected during a five-minute speech each participant gave under the supervision of two interviewers, who did not intervene during this time. The labels provided in the dataset are values of arousal and valence in the range $[-1, 1]$, sampled every 0.5 seconds. There are in total 69 samples in the dataset, each sample consisting of the five-minute-long data from each subject. In the original dataset, 41 samples are used as train set, 14 as validation set, and 14 as test set. Since annotations are not provided for the test set, as the dataset comes from a challenge, we randomly pick 4 samples from the validation set and 6 from the train set to form a new test set consisting of 10 samples. In summary, the new train set has 35 samples, the new validation set has 10 samples, and the new test set has 10 samples. More details about the ULM-TSST dataset are presented in Appendix A.3.

5.4.1.3 Input Features

We employ audio, video, and physiological signals as input modalities, using the features provided in ULM-TSST. The provided features are aligned with the annotations; that is, features are extracted every 0.5 seconds. ULM-TSST provides several types of features for the different modalities, and from them, we chose the ones listed below. Regarding the audio modality, we use the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features [64]. For the video modality, we use facial Action Unit (AU) intensity. For the physiological signals, we use Electrocardiogram (ECG), Respiration (RESP), and Beats per Minute (BPM) signals that were downsampled to 2Hz and smoothed with a Savitzky-Golay filter, so there is a 3-value feature every 0.5 seconds consisting of the concatenation of those three physiological signal values. More details about the features in the ULM-TSST dataset can be found in Appendix A.3.

We select the described features from other ones included in the dataset based on experiments run with the baseline model, which was provided by the authors of the ULM-TSST dataset, selecting the features that lead to a good performance.

5.4.1.4 Model Hyperparameters and Training

We use the Ray Tune Framework [122] to optimize different hyperparameters of our model on the validation set. The selected hyperparameters can be found in Table 5.1. Additionally to those hyperparameters, we use

Module	Hyperparameters	Activation Function
TCN	Layers: 6, Kernel size: 9, Channels: 64	ReLU
Transformer Encoder	FFN size: 256, Heads 2, Layers 2	GELU
Transformer Decoder	FFN size: 256, Heads 1, Layers 1	GELU
ERN	Layers: 1, Size: 32	ReLU

Table 5.1 – Model hyperparameters used during the experiments

a model dimension of $d_{\text{model}} = 64$. Also, the bidirectional attention mask for the input of the Transformer encoder has a *mask_length* of 50 seconds, which is equivalent to 100 time-steps.

During training, we segment each 5-minute sample into smaller overlapping samples, as suggested by Huang et al. [93] and done by other authors when working with long sequences for time-continuous emotion prediction [38, 42, 143]. According to Huang, using the overlapping segments can be seen as a way of data augmentation, multiplying the number of samples and helping the model convergence. In our case, since we use a Transformer-based approach, segmenting the samples into shorter segments also helps to perform the training more efficiently since the computational complexity of the Transformer grows quadratically with respect to the input length. Searching across different options, we found that segments of 125 seconds (250 time-steps) with a hop size of 25 seconds (50 time-steps) work well in our experimental protocol.

The model is trained for a maximum of 100 epochs, starting with a learning rate of 0.0001 and halving it if the metric does not improve for five epochs on the validation set. The training is stopped if there is no improvement in the metric for 15 epochs. We use Adam optimizer with $B_1 = 0.9$ and $B_2 = 0.999$, a dropout rate of 0.2 throughout all the model, and a batch size of 64.

Loss Function

We use CCC as the loss function for training, in line with other systems for time-continuous emotion recognition [42, 87, 128, 143]. Specifically,

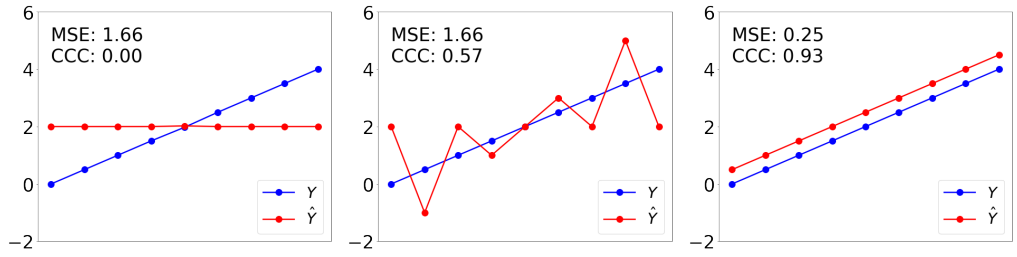


Figure 5.9 – Three examples of using CCC and MSE to measure the similarity of predicted values $\hat{Y}$ to ground-truth values Y.

using Expression 5.14, the loss is formulated as follows:

$$\begin{aligned} \mathcal{L} &= 1 - \text{CCC} \\ &= 1 - \frac{2\rho\sigma_{\hat{Y}}\sigma_Y}{\sigma_{\hat{Y}}^2 + \sigma_Y^2 + (\mu_{\hat{Y}} - \mu_Y)^2}. \end{aligned} \tag{5.16}$$

Recall that in this expression, ρ is the Pearson correlation coefficient between the predicted values $\hat{Y}$ and the ground-truth values Y. σ and μ denote the standard deviation and the mean, respectively, of the predicted or the ground-truth values, as indicated by their subindices.

We also considered using as loss the Mean Square Error (MSE) between the predicted and ground-truth values, but we preferred the CCC. While the MSE only indicates how far are the predicted values from the ground truth, the CCC value gives a measure of how far are the predicted values from the ground truth and also how correlated those values are. Figure 5.9 illustrates this, showing three examples where predicted values $\hat{Y}$ are compared to ground-truth values Y. In the examples from Figures 5.9a and 5.9b, the MSE is the same, showing that this metric fails to capture the correlation between $\hat{Y}$ and Y in Figure 5.9b. On the other hand, the CCC is better in Figure 5.9b, showing that this metric captures the correlation. We argue that the result from Figure 5.9b is preferable to the result from Figure 5.9a, and this is better indicated by the CCC value. Moreover, the CCC value not only indicates correlation, but also captures the difference in distance between $\hat{Y}$ and Y. This is shown in Figure 5.9c, where $\hat{Y}$ and Y are correlated with coefficient $\rho = 1$, but the CCC is less than one indicating the displacement of $\hat{Y}$ with respect to Y.

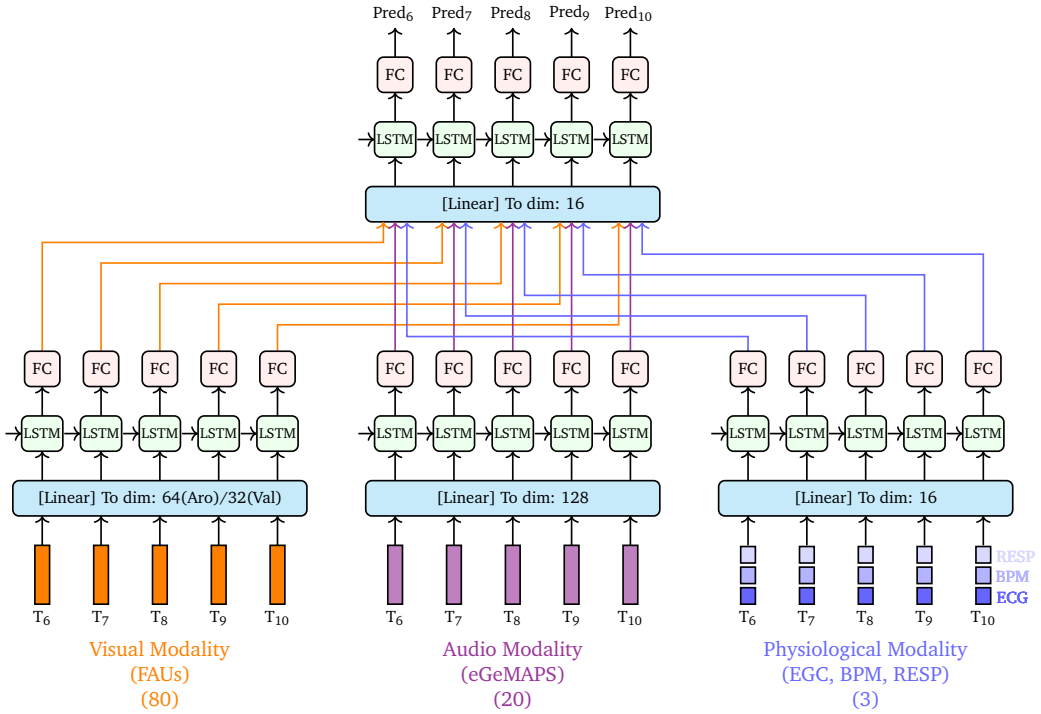


Figure 5.10 – Baseline approach Nº 1: Late-Fusion with LSTM networks.

5.4.2 Results

For each experiment, 30 results are obtained by training the model with 30 different initialization seeds, reporting the average of those results. We use a two-sided t-test with a threshold of p-value < 0.05 to assert that a result is statistically significantly different than another, using the Holm-Bonferroni correction method to account for the fact that multiple comparisons are being done.

5.4.2.1 Comparison with the Baseline

We start by comparing the results of our approach with the baseline model developed for the Muse 2022 Challenge [42], which is depicted in Figure 5.10. To test this baseline approach, we use the provided code [1], so it can be evaluated with the same features that we employ and using the same partition of train, validation and test sets of the ULM-TSST dataset that we use. This model is based on LSTM networks, using late fusion to aggregate the different modalities. This is done in two steps. In the first

1. https://github.com/EIHW/MuSe2022

Approach	AROUSAL		VALENCE	
	RMSE↓	CCC↑	RMSE↓	CCC↑
Late-fusion with LSTM [42]	0.3046 (0.020)	0.2702 (0.026)	**0.1585** (0.016)	0.1273 (0.053)
MMTE + AMMTD (ours)	**0.2948** (0.013)	**0.3578** (0.033)	0.1796 (0.011)	**0.1502** (0.027)

Table 5.2 – Comparison of our results with the baseline. The best result is indicated in bold, and we show the standard deviation in parentheses. In all cases, the differences between both approaches are statistically significantly different. The symbols (↓) and (↑) indicate that a lower and a higher score are desirable, respectively.

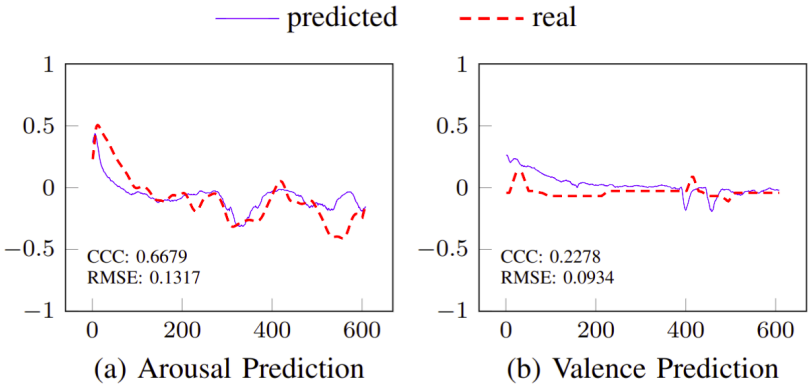


Figure 5.11 – Example of an output of our model compared with the ground-truth, when predicting arousal (a) and valence (b) for the same sample.

step, a different LSTM is trained for each modality to predict the emotion value. In the second step, the decisions of each single-modality model are concatenated and used as input for a meta regressor, also based in a LSTM network. Linear layers are used throughout the model to change the dimensions of the representations, as shown in Figure 5.10.

Table 5.2 shows the results of our model along with the results of the baseline. In all cases, the differences between the results are statistically significantly different. This table shows that for all metrics except for valence RMSE, our model obtains better results than this baseline. These results are in line with Expected Result 5.1, showing that most of the time our attention-based approach leads to better results than a no-attention-based architecture. An example of an output of our model and ground-truth values is depicted in Figure 5.11, which shows that real valence values tend to be flat and have less variability than the arousal values, which we noted is a common occurrence in our test set. We hypothesize that the baseline does

	AROUSAL CCC	VALENCE CCC
He et al. [87]	**0.6818**	**0.6841**
Liu et al. [128]	0.6689	0.6803
Park et al. [143]	0.6196	0.6351
Ours	0.3453	0.5821

Table 5.3 – Comparison of our results with other results of the Muse 2022 Challenge.

better in valence RMSE because the simpler architecture of the baseline is good enough to produce flat sequences of valence values that are close enough to the also flat ground truth. On the other hand, the baseline approach fails to predict the small changes in the valence values, penalizing the CCC score, while our model does a better job in this case.

To further compare our approach with other contributions, Table 5.3 shows the results of the top 3 entries in the Muse 2022 Challenge, in which they addressed continuous emotion recognition using the ULM-TSST dataset. In this case, the results of our approach were obtained using the official train-validation-test partition of the challenge, and submitting the predictions to the challenge web server, since the labels of the test set are not provided. The participants of the challenge typically search for the best feature combinations, sometimes using different models for each emotion dimension. For example, He et al. [87], who are the winners of the challenge, use a combination of five different types of features extracted from four modalities to predict arousal, processing those features with a Transformer-based model. For valence, they use four different features extracted from three modalities and an LSTM-based early fusion model. Different from this, our goal is to find a general architecture that works for arousal and valence, so we use the same architecture to predict both emotion dimensions. In addition, instead of tuning our approach to find the best feature combination, we found in early tests a combination of features (one type of feature per modality) that worked well, and then we improved our architecture keeping the selected features fixed. Taking into account those remarks, Table 5.3 shows that our model obtains good results for valence, although there is room for improvement regarding arousal.

5.4.2.2 Testing the Effectiveness of the AMMTD decoder

To test the importance of the AMMTD decoder, we conduct an ablation study by replacing this module with two different architectures. The idea

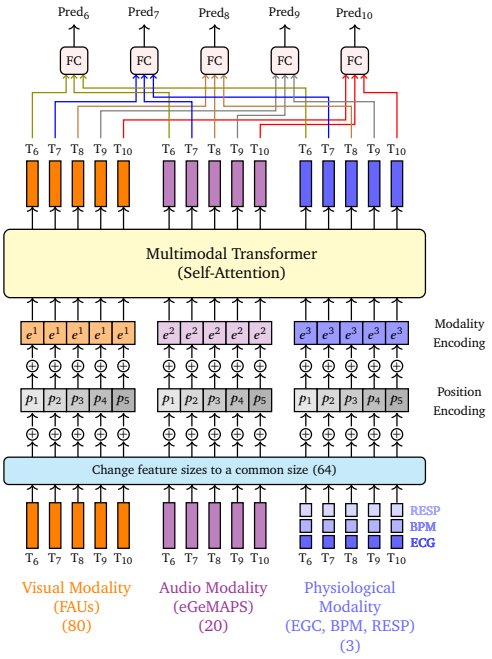


Figure 5.12 – Alternative approach Nº 1: MMTE + FCN

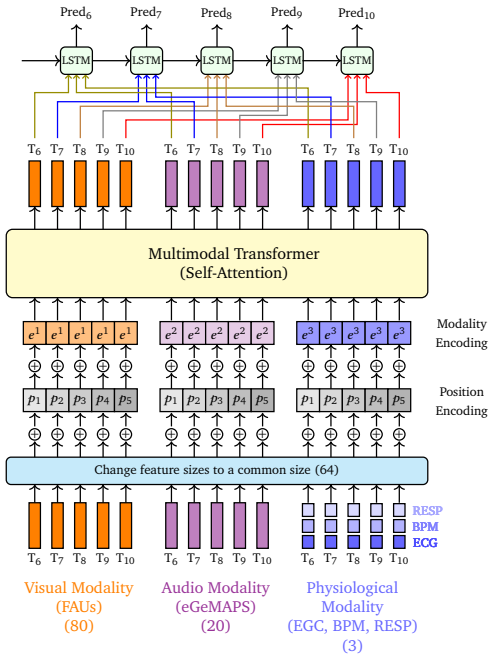


Figure 5.13 – Alternative approach Nº 2: MMTE + LSTM

is to compare other alternatives to predict the emotion values from the representations given by the MMTE encoder. The alternative architectures are described below.

Alternative Approach Nº 1: MMTE + FCN: This approach corresponds to a model where instead of using the AMMTD, the representations from the MMTE are directly processed with a FCN, as indicated in Figure 5.12. Specifically, when predicting the emotion value at time t, the representations $r_t^m \in \mathbb{R}^{d_{\text{model}}}$ of each modality m corresponding to time t are concatenated to form the vector $R_t = [r_t^1; \ldots; r_t^M] \in \mathbb{R}^{d_{\text{model}} \cdot M}$, where (;) denotes concatenation, M is the number of modalities, and d_{model} is the size of the representations given by the MMTE module. Then, the vector R_t is used as input for the FCN, and the whole architecture is trained in an end-to-end fashion.

Alternative Approach Nº 2: MMTE + LSTM: This approach uses an LSTM to process the representations given by the MMTE. An LSTM is used because it is capable of modeling the temporal relations of the sequence of representations given by the MMTE. The input for this LSTM is the sequence $[R_1, \ldots, R_T]$, where T is the sequence length and R_t is the concatenated

Approach	AROUSAL		VALENCE	
	RMSE↓	CCC↑	RMSE↓	CCC↑
MMTE+FCN	0.3238 (0.023)	0.1388 (0.068)	0.1850 (0.017)	0.1221 (0.031)
MMTE+LSTM	0.3189 (0.024)	0.1387 (0.058)	0.1842 (0.065)	0.0435 (0.064)
MMTE + AMMTD (ours)	**0.2948** (0.013)†	**0.3578** (0.033)†	**0.1796** (0.011)	**0.1502** (0.027)†

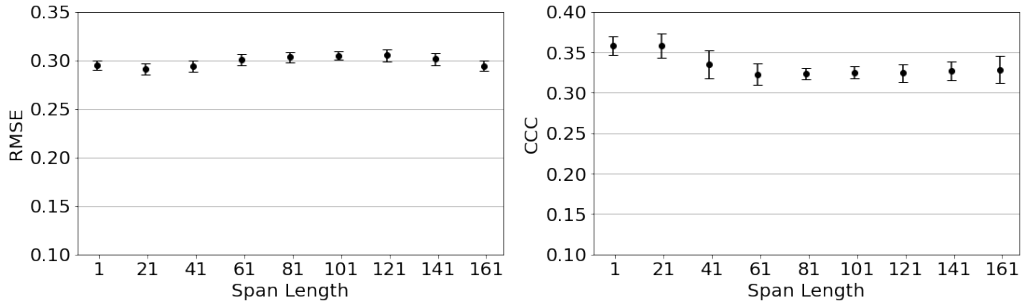
Table 5.4 – Comparison of using an LSTM and a FCN as alternatives to the AMMTD module to predict emotion values from the MMTE representations. The best result is indicated in bold, with the standard deviation in parentheses. The symbol (†) indicates if the result of our approach is statistically significantly different than the alternative approaches. The symbols (↓) and (↑) indicate that a lower and a higher score are desirable, respectively.

representations given by the MMTE, as presented in the explanation of Approach Nº 1. We used a grid search to tune the size of the LSTM, selecting an LSTM with 4 layers and a hidden dimension of 32. Figure 5.13 shows a depiction of this approach.

Table 5.4 shows the results of the alternative approaches and of our approach. These results show that the AMMTD module leads to better performance in all metrics. The results of our method are statistically significantly different from the alternative approaches in all metrics except for valence RMSE, where although our approach outperforms both baselines, the improvement is not statistically significant. These results demonstrate the effectiveness of our ideas of using cross-attention and autoregression, both implemented in the AMMTD module, to predict time-continuous values of arousal and valence. Moreover, this validates the Expected Result 5.1, showing that using attention to aggregate the information from the different modalities has superior performance than other alternatives of information fusion.

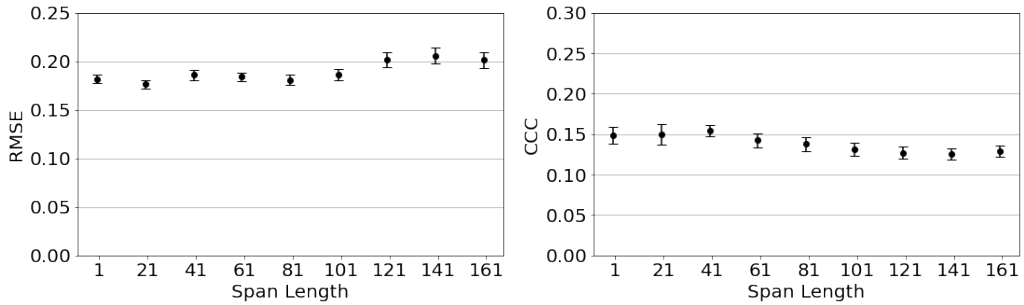
5.4.2.3 Influence of the Span of the Cross-Attention Mechanism

We define the span of the cross-attention mechanism as the number of time-steps that the MHCA module inside the AMMTD decoder attends from the representations of each modality. For example, if the span is 11, it means that the MHCA module is attending the current representation time-step r_t^m of each modality, plus the five previous and subsequent time-steps, i.e. the module is attending the sequence $[r_{t-5}^m \dots r_{t+5}^m]$ of each modality m. According to Expected Result 5.2, we expect that varying the attention span



(a) Span length vs. arousal RMSE

(b) Span length vs. arousal CCC



(c) Span length vs. valence RMSE

(d) Span length vs. valence CCC

Figure 5.14 – Plots of cross attention span length vs. RMSE and CCC when predicting arousal and valence. Bars indicate confidence intervals for a Student's t distribution, with a 95% confidence level.

will influence the model performance.

We check Expected Result 5.2 by predicting values of arousal and valence with different cross-attention span lengths. Thirty models were trained to predict arousal with different initialization seeds, and we present the average of those thirty results. Similarly, thirty models were trained to predict valence using different initialization seeds, and their results were averaged. Figure 5.14 shows these results, and as we expected, varying the cross-attention span length changes the performance of the model in terms of RMSE and CCC, confirming Expected Result 5.2. However, it is interesting that the model tends to perform better with shorter cross-attention spans. We believe this is the case because the representations generated by the encoder already incorporate temporal information, so the decoder does not need to look *far* in time. With this, the decoder can concentrate on weighting and aggregating the multimodal information, rather than modeling the temporal dependencies. Figure 5.14 shows that the best cross-attention spans are between 1 and 21, with no significant

difference between them. Thus, in our model, we used a cross-attention span of 1 since it is more computationally efficient, as the decoder has fewer representations to process.

5.5 Conclusions

This chapter presented the third contribution of this thesis, which is a method for continuous emotion recognition from multimodal inputs. Our attention-based solution addresses the challenges described in Section 5.1.3: Challenge 5.1, how to model temporal dependencies; Challenge 5.2, how to aggregate multimodal information; and Challenge 5.3, how to take into account past predictions.

To address Challenge 5.1, we design a multimodal Transformer encoder, using the attention mechanisms inside the Transformer to model the temporal relations in the input modalities. To address Challenge 5.2, we use the cross-modal attention from a Transformer decoder, thus aggregating information in a way that more weight is given to the more important modalities. Finally, to address Challenge 5.3, we employ an autoregressive approach.

The different experiments showed that our attention-based approach outperforms a baseline based on late fusion and recurrence. Moreover, an ablation study revealed that processing the representations generated by the encoder using the AMMTD increases the performance compared to using other solutions such as a FCN or a LSTM network.

In this chapter, we used multimodal signals that a smart environment may provide in order to perform continuous emotion recognition. Besides of having complementary information in the different modalities, there is an additional advantage in using multimodal information: if a modality is missing, emotions could still be inferred from the remaining ones. We explore this idea in the following chapter, where we address the task of accommodating missing modalities in continuous multimodal emotion recognition.

CHAPTER 6

ACCOMMODATING MISSING MODALITIES FOR EMOTION RECOGNITION

If we want to effectively monitor the mental well-being of a frail person in a smart environment, it is desirable to monitor the emotional state of those people by taking advantage of the variety of sensors that might be present in such an environment. To achieve this goal, there are systems that are capable of processing multimodal inputs in order to recognize emotions, like the system we presented in the previous chapter. However, in real-world scenarios, some modalities may become unavailable. For example, a sensor might run out of battery, it might have communication issues, or maybe the user deactivates a sensor for privacy concerns. These stations may cause the system not to work if it is not flexible enough to deal with these circumstances. Therefore, if we envision the deployment of emotion recognition systems in real-world settings, those systems should be capable of accommodating missing modalities.

This chapter presents a method to recognize emotions from multimodal inputs capable of handling missing modalities. The chapter gives a concrete

definition of the problem being addressed in Section 6.1, reviews some related contributions in Section 6.2, describes our contribution to handle missing modalities for emotion recognition in Section 6.3, and shows the experimental results obtained when evaluating our ideas in Section 6.4.

6.1 Problem Definition, Motivations and Challenges

In this Chapter, our goal is similar to the one that we had for Chapter 5, which is described in Section 5.1. That is, we want to perform multimodal continuous emotion recognition. For this chapter though, we address the case that modalities are missing during inference. Thus, to be more specific, in this chapter, we design a system capable of performing multimodal value-continuous time-continuous recognition of arousal or valence, even in the situation when there are missing modalities.

Our motivations behind this task come from the real-world scenario that, although all modalities may be available during training, there might be modalities that are unavailable when performing the inference. For example, if the input consists of video, audio, and physiological signals, as it is our case, the camera's field of view may become obstructed, the microphone may be too far away, the physiological sensor may be on a wearable device that is not currently worn, or simply any of the mentioned sensors may be disconnected by the user for privacy concerns.

Similar to Chapter 5, we refer to the problem of multimodal value-continuous time-continuous recognition of arousal or valence simply as *multimodal continuous emotion recognition*, denoting with the word *continuous* value-continuous and time-continuous, and with the word *emotion* either arousal or valence.

6.1.1 Challenges

In this chapter, a missing modality can be understood as a modality that is absent. That is, the input is not replaced by any value (e.g. a null value or zeros), but rather is completely unavailable. In this case, the model should use as inputs the remaining modalities. We can imagine this case of an absent modality when the system has been designed assuming that visual, audio, and physiological information is going to be present, but for example,

the final user prefers not installing cameras to gather the visual information or decides to disable the microphone when someone is visiting.

Also, in this chapter we consider that a missing modality could be a modality with invalid data, that is, the data from the modality is spurious or noisy, such that it is not feasible to reliably perform emotion recognition with these data. In this case, we assume that there is a method that identifies that a modality is not valid, and the model should ignore it. One way to check the validity of a modality is by checking the correlation between the different modalities, as done by Mittal et al. [136]. In any case, in this chapter, we do not study methods to identify invalid modalities, and we simply assume that a perfect method exists. In real life, an invalid modality might arise, for example, when there are communication errors with the sensors, such that the system keeps receiving data, but these data are not valid.

For simplicity, we refer to a modality that is absent or invalid as a missing modality. With this, one challenge addressed in this chapter is the following:

Challenge 6.1. *Developing an architecture to perform multimodal emotion recognition capable of accommodating missing modalities, such that the architecture is flexible to work with fewer modalities than originally intended, or capable of ignoring modalities identified as not valid.*

Even if an architecture is flexible enough to accommodate missing modalities when performing emotion recognition, it might have its accuracy reduced when a modality is missing. From this, the following challenge arises:

Challenge 6.2. *How to improve the accuracy of a model designed for multimodal emotion recognition capable of accommodation missing modalities, in the case that the accuracy of the model is reduced when a modality is missing.*

To address these challenges, we assume that all modalities have an equal probability of being missing, and we also assume that only one modality will be missing at a time. Although these assumptions may not hold in real life, we use them as a starting point in our approach, leaving the impact of relaxing these assumptions as future work.

6.2 Related Work on Handling Missing Modalities for Emotion Recognition

Early contributions that address the issue of missing modalities in multimodal approaches include the work of Kapoor and Picard [99], which proposes an approach based on a Mixture of Gaussian Processes to fuse the information from multiple modalities in a scenario where there might be modalities missing. Kapoor and Picard [99] demonstrate that their method can handle missing modalities better than simply stacking the observations of all modalities to form a vector used as input to a classifier.

For the rest of this literature review, we focus on Deep Learning (DL) approaches. Zhao et al. [222] identify three main types of approaches to handle missing modalities:

— Learning a joint representation from the different modalities, such that if a modality is missing at test time, the remaining ones can still generate this joint representation.

— Using generative methods to generate the missing modalities from the available ones.

— Ablating the inputs at training time to mimic the case that a modality is missing.

The following subsections examine these approaches in detail, reviewing relevant literature that showcases how these techniques are used to handle missing modalities. Specifically, contributions that learn joint representations are reviewed in Section 6.2.1, contributions that use generative methods are reviewed in Section 6.2.2, and contributions that ablate the inputs during training are reviewed in Section 6.2.3.

6.2.1 Learning Joint Representations

The idea behind this technique is to learn joint representations from the different modalities, such that these representations can be generated even when a modality is absent at test time. The key fact is that during training, the model learns to generate joint representations that capture semantic information from all the modalities. Then, during inference, the model generates these representations using only the available modalities as input. To exemplify this, we now describe in detail the works of Aguilar et al. [3] in Section 6.2.1.1 and Pham et al. [147] in Section 6.2.1.2.

6.2.1.1 Multi-view Approach for Missing Modalities

In their work, Aguilar et al. [3] address the task of emotion recognition from lexical (i.e. transcripts) and acoustic information. They argue that although during training it could be possible to access both modalities, at test time it might be more difficult to have access to the transcript. Thus, they develop a system to combine acoustic and lexical modalities during training, while for inference the system does not require lexical inputs.

For this, they induce semantic information from a multimodal model into an acoustic-only model using a multi-view approach with contrastive learning. In other words, they consider the acoustic information as one view and the multimodal information as a second view from the same input, hence their representations should be similar. Specifically, they build a model that uses only acoustic inputs, and a multimodal model that takes acoustic and lexical inputs. During training, each model is taught to predict emotion while contrastive loss is used to tie the representations generated by both models. With this approach, the authors claim that information is shared between the models, and therefore at test time the acoustic model is capable of predicting emotions without the need of lexical information, having taken advantage of this information at training time. In fact, the accuracy of predicting emotions with the acoustic model using this approach is around 10% higher than using the same model but trained only with acoustic inputs, without using the multimodal model and the contrastive loss.

The work of Aguilar et al. [3] shows that it is possible to learn a joint representation that incorporates the information from the available modalities at train time, and then generate this representation at test time using only one modality. However, one drawback of this approach is that if all modalities are present during testing, the model cannot use all of them and take advantage of this situation, since it only accepts one type of modality as input.

6.2.1.2 Learning Joint Representations with Cyclic Translations

Pham et al. [147] develop a model to predict emotions that is trained using text, visual and audio information, but uses only text to make the predictions. For this, they use a sequence-to-sequence approach to perform translation between a source modality (text) and a target modality (audio or visual), arguing that this method provides a way of learning a joint representation that uses only one modality, the source modality, as input.

Their idea is to encode the information from the text modality into an intermediate representation, and then generate the other modalities from this intermediate representation. This intermediate representation captures joint information from the source and target modalities, and thus the authors refer to it as a joint representation. To ensure that this joint representation captures most of the information from all modalities, they use a cycle consistency loss, meaning that they translate back from the generated modality to the source modality. The joint representation is used to perform the emotion prediction, and during training, the translation loss, the cycle loss, and the prediction loss are minimized together.

The key aspect of this approach is that once the model has been trained, the joint representation can be generated using only the source modality, which in this case is the text modality. Their model is capable of performing better than other works that address the same problem, even using only the text modality as input during test time, while other approaches use text, visual and audio modalities as input.

Pham et al.'s approach [147] shows how to generate a joint representation using a single modality as input, while incorporating the information from other modalities during training. However, like the work of Aguilar et al. [3], which was reviewed in the previous section, it cannot take advantage of other modalities if they are present at test time. Moreover, as these approaches rely on a single modality at test time, if this single modality is missing, the system will stop working.

6.2.2 Using Generative Methods

In this type of approach, the idea is that if a modality is missing, it will be generated from the remaining modalities using a generative method. A way to do this is to use linear transformations to generate the missing features from the available ones, like in Mittal et al. [136], which is reviewed in Section 6.2.2.1; use an encoder-decoder model like Tsai et al. [191], which is reviewed in Section 6.2.2.2; or use adversarial networks like Cai et al. [31], which is reviewed in Section 6.2.2.3.

6.2.2.1 Generating Proxy Features for Missing Modalities with Linear Functions

In their work, Mittal et al. [136] address the task of emotion recognition from facial, textual, and speech inputs. Their approach checks for ineffective modalities (for example a noisy modality), and if a modality is identified as

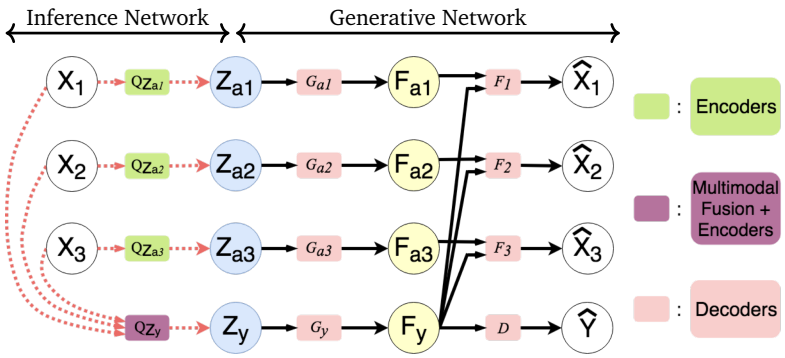


Figure 6.1 – Tsai et al.'s approach (figure from Tsai et al. [191].)

ineffective, its features are replaced by proxy features generated from the remaining effective modalities.

The first step in their approach is to verify if a modality is effective or ineffective. To do this, they argue that when emotions are predicted correctly, there exists a correlation between the input modalities. Thus, they check the correlation between pairs of modalities to identify a modality that may be uncorrelated with the others, and this modality is identified as ineffective. If a modality is identified as ineffective, a proxy feature vector is generated to replace it using a linear transformation on the features of another modality. In reality, there is a non-linear relation between modalities, but they demonstrate that by relaxing this constraint, a linear transformation can be found that approximates the ineffective modality, and an approximate feature (i.e. a proxy feature) can be obtained from the available modalities features using a linear algorithm.

An advantage of Mittal et al.'s approach [136] is that it is capable of identifying and reconstructing the representations of modalities. However, there is no guarantee that the generated representation accurately resembles the missing one, which may downgrade the performance of the model in case the proxy representation is far from the real one.

6.2.2.2 Using an Encoder-Decoder Model to Generate Missing Modalities

Tsai et al. [191] develop an architecture to generate multimodal representations that can be used for different downstream tasks, factorizing these representations into multimodal discriminative factors and modality-specific generative factors. Their method is depicted in Figure 6.1. In their approach, the model produces those factors rather than generating the representations

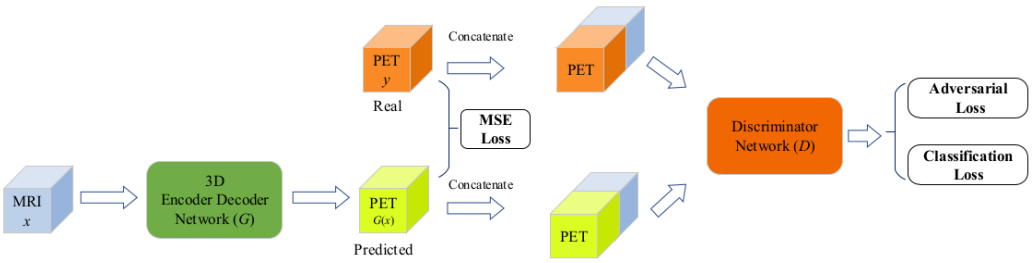


Figure 6.2 – Cai et al.'s approach (figure from Cai et al. [31].)

directly.

To do this, they employ a multimodal fusion model that produces a representation that is used to perform the prediction, this representation being the multimodal discriminative factor. They also employ different networks to generate individual representations from each modality and reconstruct the respective modality from the corresponding representation. These individual representations are the modality-specific generative factors.

At test time, if a modality is missing, the model is capable of reconstructing the missing modality from the remaining ones, being also able to produce the multimodal discriminative factor that is used for the prediction. Therefore, the model is able to perform the prediction even with the missing modality.

The results shown by Tsai et al. [191] demonstrate that in the case of missing modalities, using their factorized representation approach leads to better results compared to using a purely generative or a purely discriminative model. However, a disadvantage of their approach is that different models need to be trained to deal with different missing modalities. In other words, if there are modalities A, B, and C, a model needs to be trained in case modality A is missing, another one in case modality B is missing, and a third one in case modality C is missing.

6.2.2.3 Using Adversarial Networks

In their work, Cai et al. [31] address the problem of processing medical images for disease diagnosis. Specifically, they use magnetic resonance images and tomographies as inputs during training, with the aim of using only the magnetic resonance images during inference. For this, they use Generative Adversarial Networks (GANs) to generate the missing tomographies.

In their approach, which is depicted in Figure 6.2, they use an encoder-decoder network as generator, while their discriminator has the particularity that it is not only in charge of distinguishing real and fake input images, but also performs the target classification task. To train their model, the loss is based on the distance between the generated and real tomographies, plus the adversarial loss and plus the classification loss.

The results presented in Cai et al.'s paper [31] show that their approach can perform better in terms of accuracy than using only magnetic resonance or only tomographies, showing that the generated missing modalities help the model to achieve better results. Nevertheless, their model is not capable of accepting more than one modality as input, and thus, if all modalities are present during inference, their approach cannot take advantage of this situation.

6.2.3 Ablating Inputs at Training Time

The main idea in this approach is to ablate or eliminate different modalities at training time, such that the model learns to handle the case of missing modalities during evaluation. We now review some contributions that use this technique to handle missing modalities.

6.2.3.1 Random Dropping of Modalities

The work of Neverova et al. [140] addresses the task of multimodal gesture recognition that is robust to missing modalities. To do this, they develop a model that first processes each modality individually and then fuses the information from each modality using non-linear fusion shared layers, based on Fully-Connected Network (FCN).

Their fusion shared layers are FCNs carefully designed such that they can work even when some modalities are absent. In fact, the weight matrix of the shared layers is structured in blocks, in such a way that each diagonal block models single-modality dependencies, while the off-diagonal blocks model cross-modal dependencies. This allows to pre-train the complete model one modality at a time, initializing this way the diagonal blocks. Once this pre-training is concluded, the model could be potentially trained with all the modalities, but instead, the authors propose to drop modalities randomly, making the model robust to missing modalities at test time.

Neverova et al.'s work [140] demonstrates the potential of dropping modalities during training to make a model more robust to missing modali-

ties. However, their approach requires structuring the weights of a FCN in blocks to make it suitable to handle missing modalities. Contrary to this method, a Transformer-based approach may be seen as a more flexible technique to deal with missing modalities because there is no need to change its architecture when a modality is missing. In fact, in this case, the attention mechanisms in the Transformer will simply pay attention to the remaining modalities. The following subsection presents some works that address the problem of missing modalities using a Transformer-based approach.

6.2.3.2 Transformer-Based Approaches

Using Transformer-based approaches to process multimodal information has the advantage that if a modality is missing, the attention mechanisms of the Transformers can ignore the missing modalities and only attend to the remaining ones, thus being a solution well adapted for this situation.

One example of such an approach is the work of Parthasarathy and Sundaram [144], where a cross-modal Transformer [190] is used to model inter-modality interactions for the task of audiovisual emotion recognition (see Section 5.2.2.3 for more information about cross-modal Transformers). Two cross-modal Transformers are used: one to incorporate information from the visual modality into the audio modality, and one to incorporate information from the audio modality into the visual modality. The final representation is the addition of the output representations of the two cross-modal Transformers, plus the original audio and video features. The use of addition to obtain the final representation allows the model to work even if a modality is missing. Specifically, the authors address the case where the visual modality is missing. For this, during training, they randomly replace with zeros some selected frames.

Another example is the work of Goncalves and Busso [79], which also uses a cross-modal Transformer for audiovisual emotion recognition. In their case, in addition to the cross-modal Transformer, single-modality Transformers are also used, and the final training loss is the weighted sum of the losses of each of the Transformers. To make the model robust to missing modalities, during training the audio or visual features are replaced with zeros with a certain probability.

The contributions reviewed here show that Transformer-based approaches can be used to handle missing modalities, and they can be made more robust to this situation by dropping modalities during training. However, a disadvantage of using a cross-modal Transformer is that a single cross-modal

Transformer only accepts a pair of modalities, incorporating information from one modality into the other (and not the other way). Therefore, expanding the presented approaches to use more modalities is not straightforward.

To alleviate the situation described in the previous paragraph, a Multimodal Transformer [73] can be employed. An example of this approach can be found in the work of Ma et al. [130], where they process inputs consisting of images and text using a Multimodal Transformer, with the input formed by creating a sequence with the concatenation of the image and text features. During training, they sometimes hinder the attention between modalities, essentially converting the Multimodal Transformer into two single modality Transformers, thus now the model has two outputs. The training loss is the sum of the loss when all modalities are present and the two losses that are obtained when the attention between modalities is masked.

6.2.4 Discussion

This review has presented different types of approaches for handling missing modalities in a multimodal task. One approach is learning joint representations that capture semantic information from the different modalities, but can be generated only with a single modality. Although the reviewed contributions show that this approach is suitable to handle the situation of having only one modality available during inference, the drawback is that this approach cannot take advantage of using more than one modality in the case that all modalities are available at test time.

A second approach is to generate the missing modalities from the remaining ones. This can be done using methods like GANs, encoder-decoder approaches, or linear functions to generate the missing modalities or proxy features for the missing modalities. In this case, the model can use all the available modalities, but there is no guarantee that the generated modality or the proxy feature will accurately resemble the missing one.

To avoid generating synthetic information, a third approach consisting of dropping modalities at training time can be used. This way, the model learns to handle the case when a modality is missing. Given its advantages, this is the technique that we adopt to accommodate missing modalities. Moreover, we use a Transformer-based approach, since its architecture is well adapted to deal with missing modalities, and allows dropping modalities, as shown by Gabeur et al. [72]. Specifically, we use the architecture introduced in

Chapter 5, composed by a Multimodal Transformer Encoder (MMTE) and a Autoregressive Multimodal Transformer Decoder (AMMTD). Moreover, instead of randomly dropping any modality during training as is typically done, we drop the modalities that we identify as the ones that are contributing the most to the prediction, forcing the model to learn to obtain information from the other modalities. In summary, our approach takes advantage of the flexibility of Transformer-based approaches to accommodate missing modalities, while employing a modality-drop strategy to better prepare the model to the case when a modality is missing.

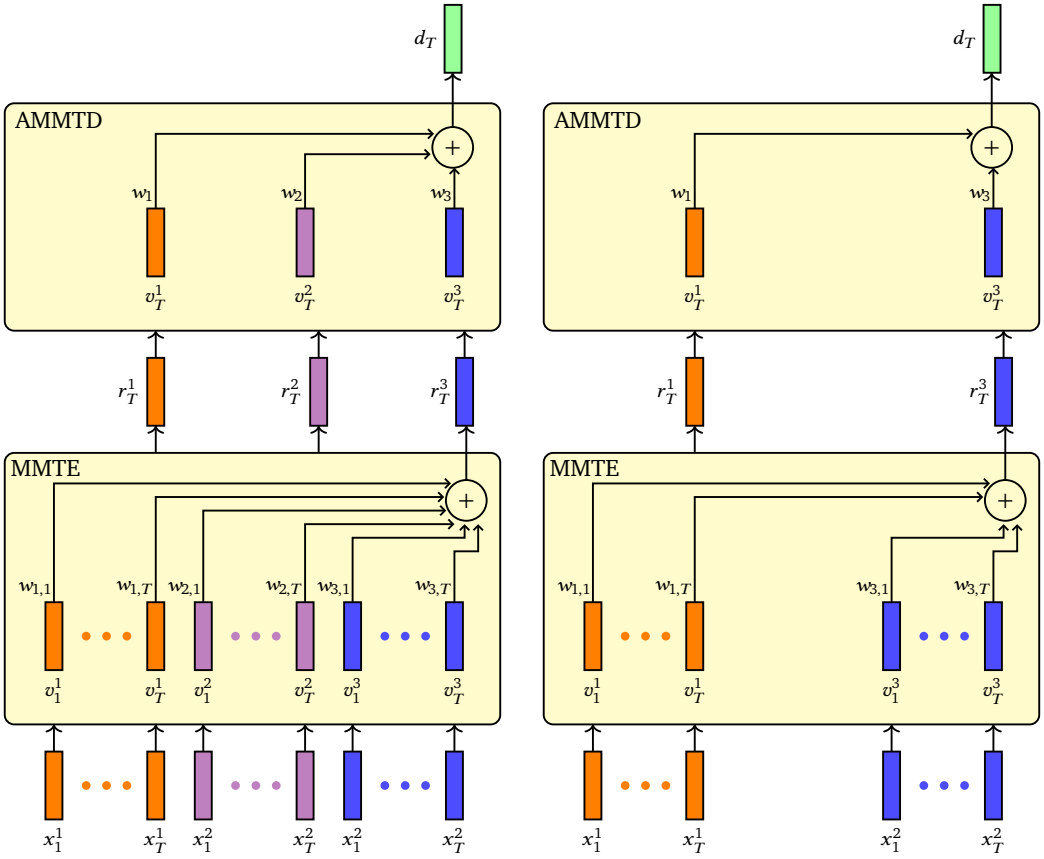
6.3 Accommodating Missing Modalities

The contribution of this chapter, which is described in this section, is to develop an approach to handle missing modalities when recognizing value-continuous time-continuous values of arousal or valence, using multimodal inputs. For simplicity, we use the term *multimodal continuous emotion recognition* to refer to recognizing value- and time-continuous values of arousal or valence from multimodal inputs.

For our approach, we use the same architecture that we developed in Chapter 5. As we did in that chapter, we use a common architecture to predict arousal and valence, but we train different models for each one of them. The architecture consists of two modules. The first module is an encoder that we call MMTE, which is in charge of generating representations from the multimodal inputs. The second module is a decoder that we call AMMTD, which uses the multimodal representations from the encoder to perform continuous emotion recognition, aggregating the multimodal information and using auto-regression to take into account past predictions. Please refer to Section 5.3 for more details about the architecture.

6.3.1 Accommodating Missing Modalities in an Attention-Based Architecture

Different from an approach where the fusion of information is done explicitly, like concatenating features for example, our model will not break in the case a modality is missing. In fact, the attention mechanisms in our Transformer-based approach can accommodate missing modalities by simply *not attending* them. We now explain this idea in detail, starting with how the MMTE behaves when a modality is missing.



(a) All modalities present.

(b) Having a missing modality.

Figure 6.3 – Depiction on how our Transformer-based architecture is flexible in the case where modalities are missing.

6.3.1.1 Handling Modalities in the MMTE

The input of the MMTE is the concatenation of the different modalities to form a single sequence, as observed in Figure 6.3a. If there are M modalities with length T, the input of the MMTE is

$$X = [x_1^1, \ldots, x_T^1, \ldots, x_1^M, \ldots, x_T^M], \tag{6.1}$$

where $x_t^m \in \mathbb{R}^{d_{\text{model}}}$ is a feature corresponding to modality m at time-step t.

To facilitate the explanation, and without the loss of generality, assume that the number of modalities is three. Moreover, let us concentrate on how the representation r_T^3 obtained from feature x_T^3 is generated, assuming that the attention module has only one head and one layer. The process of generating the representation r_T^3 with these assumptions is depicted in Figure

6.3a. Since we are using self-attention, the representation r_T^3 is obtained from query, key and value vectors coming from the input. Specifically, r_T^3 is the weighted sum of the value vectors v_t^m, with each value vector v_t^m obtained from the corresponding input feature x_t^m. Then, to obtain the representation r_T^3 we have

$$r_T^3 = w_{1,1}v_1^1 + \cdots + w_{1,T}v_T^1 + w_{2,1}v_1^2 + \cdots + w_{2,T}v_T^2 + w_{3,1}v_1^3, \cdots + w_{3,T}v_T^3, \quad (6.2)$$

where $w_{m,t}$ corresponds to the attention weights, which are obtained using the query and key vectors, Particularly, $w_{m,t}$ represents the weight (or attention) given to the input feature corresponding to modality m at time-step t.

When a modality is absent, the associated value vectors and attention weights are not generated. For example, assume that modality two is missing, which is depicted in Figure 6.3b. In this case, Expression 6.2 becomes

$$r_T^3 = w_{1,1}v_1^1 + \cdots + w_{1,T}v_T^1 + w_{3,1}v_1^3, \cdots + w_{3,T}v_T^3. \quad (6.3)$$

This shows the flexibility of the Transformer-based approaches to handle missing modalities because in this case, the model does not break. Instead, it accommodates the missing information by simply attending to the remaining information.

Note that this architecture can also handle the case where a modality instead of being completely absent, is identified as not valid. This could be the case if, for example, a sensor becomes faulty and gives incorrect values. Having identified the faulty modality, the model could ignore those values by masking the attention weights. In other words, the attention weights corresponding to the faulty modality become 0. As an example, assume that the faulty modality is modality two; then we have

$$\begin{aligned} r_T^3 &= w_{1,1}v_1^1 + \cdots + w_{1,T}v_T^1 + 0v_1^2 + \cdots + 0v_T^2 + w_{3,1}v_1^3, \cdots + w_{3,T}v_T^3 \\ &= w_{1,1}v_1^1 + \cdots + w_{1,T}v_T^1 + w_{3,1}v_1^3, \cdots + w_{3,T}v_T^3. \end{aligned} \quad (6.4)$$

6.3.1.2 Handling Modalities in the AMMTD

The way the AMMTD decoder works when a modality is absent is similar to how the MMTE module works in the same situation. Using the same simplifying assumptions as before (3 modalities, model with one layer and one head), let us study the case when the model is predicting the emotion at

time-step T. In this case, as shown in Figure 6.3a, the representations given by the MMTE that are used by the AMMTD to perform the prediction are

$$R = [r_T^1, r_T^2, r_T^3], \tag{6.5}$$

where $r_T^m \in \mathbb{R}^{d_{\text{model}}}$ is the representation corresponding to modality m at time-step T. Recall that in the AMMTD, the value vectors come from the representations given by the MMTE module. With this in mind, if the value vectors are $[v_T^1, v_T^2, v_T^3,]$, then the output for the time-step T of the AMMTD is

$$d_T = w_1 v_T^1 + w_2 v_T^2 + w_3 v_T^3, \tag{6.6}$$

where the weights w are the attention weights obtained using the representations R and the information from past predictions.

Like in the case of the MMTE, when a modality is absent, the associated vectors and attention weights are not generated. If, for example, modality 2 is absent, we have

$$d_T = w_1 v_T^1 + w_3 v_T^3. \tag{6.7}$$

Also, like for MMTE, when a modality is identified as not valid, the model could ignore this modality by using zeros as attention weights for the invalid modality. If, for example, modality 2 not valid, we have

$$d_T = w_1 v_T^1 + 0 v_T^2 + w_3 v_T^3. \tag{6.8}$$

To summarize this and the previous subsection, our complete architecture, composed of the MMTE and AMMTD modules, is capable of handling missing modalities without breaking in case a modality is completely absent. Also, our architecture can ignore a modality that has been identified as not valid. This means that our architecture is capable of addressing Challenge 6.1. However, even if our approach continues working in the case of missing modalities, its performance may be degraded. To alleviate this situation and increase the robustness of the model to missing modalities, we perform a special way of training that we call *optimized training*, which is explained in the following subsection.

6.3.1.3 Optimized Training for Missing Modalities

As seen in the previous subsections, our architecture is capable of working even if there are modalities absent. Nevertheless, the accuracy of the predictions may be negatively affected in this circumstance. For this, we

develop a way of training our model to mitigate this situation. Through this chapter, we call this way of training as *optimized training*.

The procedure for the optimized training consists in first identifying the most important modalities. To do this, the model is first trained in a standard way and then evaluated on the validation set, with every modality missing, one at a time. With this procedure, we can identify in which cases the performance is reduced more, which means that the missing modalities in those cases should be important. Specifically, if the difference in both employed metrics [1] when a particular modality is missing is statistically significant compared to when all the modalities are present, we consider that particular modality as important. We select as important all the modalities that meet this condition, even if it is more than one.

The next step after that the important modalities have been identified is to retrain the model without using those important modalities a portion of the time. Specifically, for each batch, the important modality i may be selected to be eliminated with probability $p^i_{\text{eliminate}}$, and all the modalities may be kept with probability $p_{\text{none}} = 1 - \sum_{i=1}^{n} p^i_{\text{eliminate}}$, where n is the number of important modalities.

The rationale behind this training strategy is that by hiding the important modalities, the model is forced to learn from the remaining ones, and this makes the model robust when those important modalities are not present. In addition to this, this training strategy should also improve the results when all the modalities are present, because more information will be taken from all the modalities instead of just relying on the important ones.

6.3.2 Expected Results

6.3.2.1 Expected Results with Missing Modalities, Standard Training

If our model has been trained under a standard procedure, we expect the following when a modality is missing:

Expected Result 6.1. *The model should continue working even when a modality is missing, although the performance of the model with respect to the employed metrics may be reduced.*

1. We discuss about metrics in the Experiments section. (Section 6.4.1.4).

6.3.2.2 Expected Results with Missing Modalities, Optimized Training

When a modality is missing, the quality of the results may be reduced. With our optimized training strategy we aim to mitigate this situation, and thus the following result is expected:

Expected Result 6.2. *The results obtained with a model trained with the optimized-training strategy and tested with a missing modality should be better than the results obtained with the same model, but trained in a standard manner, when tested with the same missing modality.*

6.3.2.3 Expected Results with All Modalities, Optimized Training

With our optimized training strategy, we force the model to use information from modalities that it may consider less important. We believe that using more information from these modalities should improve the overall performance. In short, the following result is expected:

Expected Result 6.3. *The results obtained with a model trained with the optimized-training strategy and tested with all modalities present should be better than the results obtained with the same model, but trained in a standard manner.*

6.4 Experiments

This section presents and discusses the experimental results when testing our approach for handling missing modalities in multimodal continuous recognition of arousal and valence, describing the experimental setup in Section 6.4.1, showing the results of testing the model with missing modalities in Section 6.4.2, and showing the results of evaluating the model trained with the optimized approach in Section 6.4.3.

6.4.1 Experimental Setup

We employ the same experimental setup used in Chapter 5, using the same model size and hyperparameters. We provide a summary of the experimental setup here, but a detailed description can be found in Section 5.4.1.

6.4.1.1 Database

We use the Ulm-Trier Social Stress Test (ULM-TSST) database, which consists of participants recorded giving a five-minute speech under a stressful situation. In addition to audio and video, physiological signals consisting of Electrocardiogram (ECG), Respiration (RESP), and Beats per Minute (BPM) are also collected. The dataset was annotated by experts, with continuous annotations every 0.5 seconds. Annotations consist of numerical values of arousal and valence in the range of $[-1, 1]$.

6.4.1.2 Input Features

We use three modalities as input for our model: audio, visual, and physiological signals. The ULM-TSST dataset provides features extracted from those modalities, and we use those provided features. For the audio modality, we use the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features [64]. For the video modality, we use facial Action Unit (AU) intensity. For the physiological signals, we use ECG, RESP, and BPM signals concatenated to a 3-value feature vector.

6.4.1.3 Loss Function

We use Concordance Correlation Coefficient (CCC) as the basis for the loss to train our model. A detailed explanation about the CCC can be found in Sections 5.4.1.1 and 5.4.1.4. This loss is formulated as

$$\mathcal{L} = 1 - \text{CCC} \tag{6.9}$$

$$\text{CCC} = \frac{2\rho\sigma_{\hat{Y}}\sigma_{Y}}{\sigma_{\hat{Y}}^2 + \sigma_{Y}^2 + (\mu_{\hat{Y}} - \mu_{Y})^2}, \tag{6.10}$$

where ρ is the Pearson correlation coefficient between the predicted values $\hat{Y}$ and the ground-truth values Y. σ and μ denote the standard deviation and the mean, respectively, of the predicted or the ground-truth values, as indicated by their subindices.

6.4.1.4 Metrics

We use the CCC, defined in Expression 6.10, as one of our metrics. Generally speaking, the CCC metric gives two pieces of information: how correlated are the predicted sequence of values with the ground truth sequence of values, and how far are the predicted values from the real ones.

The range of this metric is between -1 and 1, with 1 indicating a perfect prediction, -1 indicating a perfect reversed prediction, and 0 indicating there is no correlation between the prediction and ground truth. Additionally to the CCC metric, we use the Root Mean Square Error (RMSE) to have an idea of how distant are the predicted values from the ground truth.

When evaluating a model with these metrics, a higher CCC and a lower RMSE is desirable, which is indicated with the symbols (↑) and (↓) when we present the results.

6.4.1.5 Experimental Procedure

We independently test arousal and valence, training different models for each of them, although we should keep in mind that the architecture is the same in both cases. For each test, we obtain 30 results by training the model with 30 different initialization seeds, and we report the average of those results. A t-test is used with a threshold of p-value < 0.05 to assert statistical significance, with a two-sided t-test used to check if results are significantly different, and a one-sided t-test used to check if a result is significantly better than another.

6.4.2 Testing the Model With Missing Modalities

Our first experiment consists in testing the model when there are modalities missing. In this case, the model was trained in a standard way, that is, without using the optimized training procedure. In our experiments, a missing modality means the case where the modality is completely absent, so the input of the MMTE consists only of the features of the remaining modalities. This is similar to having a modality identified as invalid because the model could ignore it by setting the corresponding attention weights to zero, as seen in Sections 6.3.1.1 and 6.3.1.2.

With our first experiment, we try to corroborate the Expected Result 6.1, meaning that the model should work although a decrease in performance may be present in some cases. Table 6.1 presents the results of this experiment. First, we analyze the case for arousal prediction. It is clear in Table 6.1a that the difference between the metrics obtained with all the modalities present and when the audio or the physiological modalities are missing is small. In fact, there is no statistically significant difference in these cases. On the other hand, when the video modality is missing, the performance measured with CCC and RMSE drops significantly. Specifically, the CCC decreases from 0.3578 to 0.2713 and the RMSE increases from 0.2948 to

Metric	All Modalities	Missing Audio	Missing Video	Missing Physio
RMSE↓	0.2948 (0.0125)	0.2926 (0.0206)	0.3252 (0.0422)†	0.2920 (0.0146)
CCC↑	0.3578 (0.0317)	0.3589 (0.0282)	0.2713 (0.0378)†	0.3539 (0.0382)

(a) Arousal Results

Metric	All Modalities	Missing Audio	Missing Video	Missing Physio
RMSE↓	0.1796 (0.0114)	0.2533 (0.1036)†	0.2170 (0.0409)†	0.1808 (0.0122)
CCC↑	0.1502 (0.0272)	0.0738 (0.0360)†	0.1564 (0.0275)	0.1486 (0.0284)

(b) Valence Results

Table 6.1 – Results obtained with the model trained with standard training, tested with all modalities present and also with a missing modality. The standard deviation is indicated in parentheses. The symbol (†) indicates that a result obtained with a missing modality is statistically significantly different than the result obtained with all the modalities.

0.3252. These results confirm, for the case when predicting arousal, Expected Result 6.1, with our model able to continue working when a modality is missing, although performance is reduced in some cases.

When predicting valence, Table 6.1b shows that there is no significant performance degradation in terms of CCC and RMSE when physiological signals are missing. Regarding CCC, it goes from 0.1502 to 0.1486, while RMSE goes from 0.1796 to 0.1808. In both cases, the difference is not statistically significant. On the contrary, there is a significant performance drop when audio or video modalities are absent. For instance, RMSE increases from 0.1796 to 0.2533 when the audio modality is missing and increases to 0.2170 when the video modality is missing. Similar to the results obtained when predicting arousal, these results confirm the Expected Result 6.1, showing that our model continues to predict valence when a modality is missing, although with a drop in performance in some cases.

To demonstrate that our Transformer-based solution is superior to non-attention-based approaches for accommodating missing modalities, we use the alternative approaches introduced in Section 5.4.2.2, testing them with missing modalities. Specifically, we test Alternative Approach Nº 1, where a FCN is used instead of the AMMTD decoder, and Alternative Approach Nº 2, where a Long Short-Term Memory (LSTM) network is used instead of the AMMTD. Since in this case the FCN and LSTM networks are expecting as input the concatenation from the representations from the three modalities (see Figures 5.12 and 5.13), it is still necessary to generate the representation

for the missing modalities. In this case, we replace the missing modalities with zeros, which in turn will produce vectors of zeros as representations of those missing modalities. Therefore, the input of the FCN and the LSTM networks will be the concatenation of valid representations and representations filled with zeros coming from the missing modalities. It could be argued that it is not fair to compare these alternative approaches with our strategy, since they are not designed to handle missing modalities and they are been fed with vectors filled with zeros. So, it might be said that it is fairer to use an approach that for example, generates the missing modality, or its representations, from the remaining ones, and use those generated representations instead of vectors of zeros, as done in other state-of-the-art works (see Section 6.2.2). Nevertheless, our intention is to prove that an attention-based architecture is flexible enough to handle missing modalities, without requiring techniques like generating proxy representations to replace the missing inputs.

Table 6.2 shows the results of this experiment. This table shows that in the majority of cases, our approach performs better than the alternative approaches when a modality is missing. For example, when predicting valence, in the case when the audio modality is missing, our model obtains a CCC of 0.0738 while the MMTE+FCN approach obtains a CCC of 0.0159 and the MMTE+LSTM approach obtains a CCC of 0.0166. Moreover, for most of the cases, the proportion of the reduction when a modality is missing is less with our model than with the other approaches. For example, when predicting arousal with the video modality missing, with our approach the RMSE increases (worsen) 10.3%, while with MMTE+FCN it increases 15.4% and with MMTE+LSTM it increases 21.3%. In general, the results in Table 6.2 show that using the alternative approaches almost always presents degradation when a modality is missing, which does not happen with our approach which has many instances where the performance does not decrease with a missing modality. One such instance in which the performance does not decrease with our approach is when audio is missing when predicting arousal. This shows the superiority of our design that uses our AMMTD module to process the representations produced by the MMTE encoder. Moreover, these results show the robustness to missing modalities of Transformer-based approaches, confirming their adaptability to this type of situation.

	MMTE+FCN		MMTE+LSTM		MMTE+AMMTD (ours)	
	RMSE↓	%↓	RMSE↓	%↓	RMSE↓	%↓
All	0.3238 (0.0277)	–	0.3189 (0.0244)	–	**0.2948** (0.0125)	–
No Audio	0.3653 (0.0525)	12.8	0.3664 (0.0732)	14.9	**0.2926** (0.0206)	**-0.80**
No Video	0.3736 (0.0811)	15.4	0.3868 (0.0733)	21.3	**0.3252** (0.0422)	**10.3**
No Physio	0.3638 (0.0840)	12.4	0.3530 (0.0415)	10.7	**0.2920** (0.0146)	**-1.00**

(a) Arousal results, RMSE metric.

	MMTE+FCN		MMTE+LSTM		MMTE+AMMTD (ours)	
	CCC↑	%↓	CCC↑	%↓	CCC↑	%↓
All	0.1388 (0.0682)	–	0.1387 (0.0575)	–	**0.3578** (0.0317)	–
No Audio	0.0668 (0.0870)	51.9	0.0560 (0.1023)	59.6	**0.3589** (0.0282)	**-0.30**
No Video	0.0579 (0.1130)	58.3	0.0598 (0.0948)	56.9	**0.2713** (0.0378)	**24.2**
No Physio	0.0957 (0.0757)	31.0	0.0564 (0.0723)	59.4	**0.3539** (0.0382)	**1.10**

(b) Arousal results, CCC metric.

	MMTE+FCN		MMTE+LSTM		MMTE+AMMTD (ours)	
	RMSE↓	%↓	RMSE↓	%↓	RMSE↓	%↓
All	0.1850 (0.0167)	–	0.1842 (0.0653)	–	**0.1796** (0.0114)	–
No Audio	0.3180 (0.0730)	71.9	**0.1823** (0.0685)	**-1.0**	0.2533 (0.1036)	41.0
No Video	0.2919 (0.0995)	57.8	**0.2113** (0.0876)	**14.7**	0.2170 (0.0409)	20.8
No Physio	0.2700 (0.1221)	45.9	0.1868 (0.0619)	1.4	**0.1808** (0.0122)	**0.7**

(c) Valence results, RMSE metric.

	MMTE+FCN		MMTE+LSTM		MMTE+AMMTD (ours)	
	CCC↑	%↓	CCC↑	%↓	CCC↑	%↓
All	0.1221 (0.0309)	–	0.0435 (0.0640)	–	**0.1502** (0.0272)	–
No Audio	0.0159 (0.0259)	87.0	0.0166 (0.0355)	61.9	**0.0738** (0.0360)	**50.8**
No Video	0.0641 (0.0389)	47.6	0.0189 (0.0509)	56.7	**0.1564** (0.0275)	**-4.10**
No Physio	0.0854 (0.0467)	30.1	0.0236 (0.0474)	45.8	**0.1486** (0.0284)	**1.10**

(d) Valence results, CCC metric.

Table 6.2 – Comparison of using the AMMTD module with other approaches when a modality is missing. "%" indicates the percentage loss of the metric when a modality is missing compared to using all modalities. Best results are indicated in bold, and standard deviation is indicated in parentheses.

6.4.3 Optimized-Trained Model with Missing Modalities

With the results from the previous experiment, we notice that when performing emotion recognition, there are modalities that influence the outcome more than others. We call those modalities *important modalities*. However, other modalities that are not important modalities may still carry useful information for emotion recognition. Therefore, it is appealing to force the model to increase its use of non-important modalities, especially to prepare the model for the situation that a modality is missing. To do this, we use our optimized training strategy of hiding important modalities during parts of the training, forcing the model to rely more on the other modalities.

In order to train the model with the optimized training strategy, we first need to identify the most important modalities. We evaluate the model using the validation set, with a single modality missing at a time. Using these results, we select the modalities that, when missing, induce a statistically significant drop in performance measured with both the CCC and RMSE metrics. With this, when predicting arousal, we identified that the most important modality is video. Similarly, when predicting valence, we identified that the most important modalities are audio and video.

Using the optimized training strategy, described in Section 6.3.1.3, when training the model to predict arousal, the video modality is eliminated with probability $p^{\text{video}}_{\text{eliminate}} = 0.25$, and all the modalities are maintained with probability $p_{\text{none}} = 0.75$. In the same way, when training to recognize valence, the audio modality is eliminated with probability $p^{\text{audio}}_{\text{eliminate}} = 0.333$, the video modality is eliminated with probability $p^{\text{video}}_{\text{eliminate}} = 0.333$, and all modalities are kept with probability $p_{\text{none}} = 0.334$. These probabilities were found empirically by testing several configurations and keeping the best ones when evaluated on the validation set.

We use the model trained with the optimized strategy to check Expected Result 6.2, meaning that we expect that when a modality is missing, the results obtained with the model trained with the optimized strategy should be better than the ones obtained with the model trained in a standard way.

Table 6.3 presents the results of the model trained with the optimized strategy and tested when a modality is missing, compared with the results obtained with the model trained with the standard strategy and also tested when a modality is missing. The table shows that our optimized training strategy improves all the results when a modality is missing. For example, when the physiological signals are missing, CCC improves from 0.3539

	Standard	**Optimized**
Missing Audio	0.2926 (0.0206)	**0.2850** (0.0132)†
Missing Video*	0.3252 (0.0422)	**0.3249** (0.0317)
Missing Physio	0.2920 (0.0146)	**0.2878** (0.0160)

(a) Arousal results, RMSE metric (↓).

	Standard	**Optimized**
Missing Audio	0.3589 (0.0282)	**0.3644** (0.0501)
Missing Video*	0.2713 (0.0378)	**0.2984** (0.0345)†
Missing Physio	0.3539 (0.0382)	**0.3571** (0.0477)

(b) Arousal results, CCC metric (↑).

	Standard	**Optimized**
Missing Audio*	0.2533 (0.1036)	**0.2052** (0.0564)†
Missing Video*	0.2170 (0.0409)	**0.1809** (0.0175)†
Missing Physio	0.1808 (0.0122)	**0.1746** (0.0094)†

(c) Valence results, RMSE metric (↓).

	Standard	**Optimized**
Missing Audio*	0.0738 (0.0360)	**0.1170** (0.0333)†
Missing Video*	0.1564 (0.0275)	**0.1676** (0.0232)†
Missing Physio	0.1486 (0.0284)	**0.1637** (0.0164)†

(d) Valence results, CCC metric (↑).

Table 6.3 – Comparison of the results when a modality is missing using a model trained in a standard way with a model trained with the optimized strategy. An asterisk (*) indicates that the modality was identified as important. Standard deviation is indicated in parentheses. Bold font is used to indicate that the result is better than its counterpart trained in a different fashion, and if it is statistically significantly better, this is indicated with the symbol (†).

	Metric	Standard	Optimized
Arousal	RMSE↓	0.2948 (0.0125)	**0.2869** (0.0120)†
	CCC↑	0.3578 (0.0317)	**0.3703** (0.0351)
Valence	RMSE↓	0.1796 (0.0114)	**0.1739** (0.0089)†
	CCC↑	0.1502 (0.0272)	**0.1656** (0.0169)†

Table 6.4 – Results obtained with all modalities present, using a model trained in a standard way and using a model trained with the optimized strategy. Bold indicates the best result, the symbol (†) indicates that the result is statistically significantly better.

to 0.3571 when predicting arousal and from 0.1486 to 0.1637 when predicting valence. Notably, the improvement is statistically significant in all cases when the important modalities are missing, except for RMSE when predicting arousal with the video modality missing. These results confirm Expected Result 6.2, showing that our optimized training strategy works well, making our model less reliant on the important modalities and using more information from the other ones.

6.4.3.1 Optimized Training Strategy with All Modalities Present

With the optimized strategy we force the model to use more information from the non-important modalities. We believe that this information may carry important cues to recognize emotion. Therefore, forcing the model to use more of this information should improve the results not only when a modality is missing, but also when all the modalities are present. This is in fact what we expect according to Expected Result 6.3.

Table 6.4 shows a comparison between the results obtained using the model trained with the optimized strategy and the results using the model trained in a standard way, when all the modalities are present. The table shows that for both arousal and valence, and for both metrics, the model trained with the optimized strategy performs better. This confirms Expected Result 6.3, showing that the model is getting more information from the non-important modalities, and this information is contributing to improving the performance of the model.

6.5 Conclusions

This chapter presented the final contribution of this thesis, which is a model to perform multimodal continuous emotion recognition, that can accommodate missing modalities.

We demonstrated that our Transformed-based approach is flexible enough to accommodate missing modalities without any architectural modifications. The self-attention layers in the MMTE module can accommodate the missing modalities by paying attention to the remaining ones, as well as the cross-attention layers in the AMMTD will just attend to the representations generated from the available modalities. However, there are cases when the performance drops when a modality is missing. In fact, we showed that there are modalities that when absent, impact the performance more than others, indicating that the model uses more information from these *important* modalities paying less attention to the others.

To alleviate the situation described in the previous paragraph, we introduced an optimized training strategy, which consists in hiding the important modalities during training. We showed that with this optimized strategy, the performance of the model increases not only when a modality is missing, but also when all modalities are present. This performance increment was especially important when the missing modalities were important modalities. These results demonstrate that by hiding the important modalities during training, the model learns to obtain more information from the other modalities.

In our work, we used the ULM-TSST dataset, with audio, visual and physiological data. Further work is needed to see if our results generalize to other datasets and to other modality combinations. Moreover, our approach is not designed to identify the case that a certain modality is present, but it is not valid, as we assumed that there is a method that identifies an invalid modality. In fact, having invalid data is a highly-probable scenario in a real-world situation, thus addressing this problem is necessary if we envision the deployment of emotion recognition systems in real-world settings.

CHAPTER 7

CONCLUSIONS AND PERSPECTIVES

In this chapter, we conclude the manuscript by summarizing our contributions. We also examine the limitations of our work and discuss perspectives to address those limitations.

7.1 Conclusions

The goal of this thesis was to develop techniques to perform emotion recognition. Through our work to address this task, we made contributions to recognizing emotions from single physiological signals, multiple physiological signals, and multimodal inputs. We also addressed the issue of missing modalities when performing the recognition. We synthesize those contributions below.

7.1.1 Emotion Recognition from Single Physiological Signals

Our first contribution was to propose a method to recognize emotions from single physiological signals. In our approach, we used as inputs raw physiological signals, employing a Deep Learning (DL) model to extract representations from those signals. This raises the question of how to effectively process such signals. For this, we investigated a Transformer-based approach. We demonstrated that a Transformer concentrates on processing the most informative parts of an input signal.

A supervised DL approach for constructing a model for physiological signals requires labeled training data. This poses a problem since a characteristic of DL approaches is that satisfactory performances typically depend on having enough data to train the model [185], and large labeled datasets of physiological signals with enough data to train the model effectively are difficult to obtain. To overcome this issue, we proposed a self-supervised pre-training technique, using unlabeled physiological signals. This pre-training technique consisted in masking some segments of the input signal and then predicting those masked segments.

Using two different datasets and two different types of physiological signals, we experimentally showed that a model pre-trained with our technique and then fine-tuned with labeled data was less prone to overfitting and had better performance in terms of accuracy and F1-score than a model trained from scratch. We also showed that with our approach, we obtained better results than a state-of-the-art work that uses a different architecture and pre-training strategy, when both our approach and the state-of-the-art approach used the same experimental protocol.

Our results demonstrate that it is valuable to use Transformer-based solutions to process raw physiological signals. Moreover, we showed that pre-training is an adequate technique for recognizing emotions from physiological signals, making the model less prone to overfitting, which is especially important in data-constrained scenarios typically found in affective computing.

7.1.2 Emotion Recognition from Multiple Physiological Signals

Extending our first contribution, for our second contribution we proposed a method for emotion recognition from multiple physiological signals.

One of the advantages of using multiple physiological signals is that the information contained in them may be complementary, and therefore combining that information should improve the results compared to using each signal individually.

As was the case for single physiological signals, it could be a challenge to find large enough labeled datasets with all the concerned signals to effectively train a DL model. Thus, as we did for our first contribution, we employed a pre-training technique, but taking into account that acquiring datasets containing all relevant physiological signals is more challenging than obtaining multiple single-signal datasets, even if these datasets do not require labeling.

Our pre-training strategy consisted in pre-training and fine-tuning individual Transformer-based models, with each model processing a single physiological signal. Then, we used a late-fusion approach to combine the results of each individual model. This way, during pre-training, the datasets used for this phase only needed to have a single physiological signal.

Through experimental results using two different datasets, we showed that using pre-trained individual models led to better results in terms of accuracy and F1-score than using individual models trained from scratch. In addition, we showed that one of the reasons for this improvement was that the model was less prone to overfitting. This was the case even if the individual models were frozen during late-fusion training, demonstrating that the representations generated by the pre-trained individual models were robust in the sense that using them as inputs for the late-fusion model made the system less prone to overfitting. In addition, experimental results showed that combining multiple physiological signals is helpful, increasing in most of cases the performance in terms of accuracy and F1-score compared to using individual signals.

Our results using multiple physiological signals show that information in different physiological signals is complementary, and therefore it is worth using multiple signals since there is the potential for a performance increase. In addition, we showed that using pre-training is also useful in the case of combining multiple physiological signals, which as we mentioned before, is especially important in the affective computing field where datasets tend to be small.

7.1.3 Multimodal Time-Continuous Emotion Recognition

For our third contribution, we proposed a method to perform time-continuous emotion recognition using multimodal inputs. When addressing this problem, we identified three main challenges: how to model the temporal dependencies present in each modality, how to aggregate multimodal information, and how to use past emotion predictions when inferring the current emotion.

We used an encoder-decoder approach to address this task. The encoder processes the multimodal inputs and generates representations from those inputs, and the decoder uses those representations and gives as an output the recognized emotion. In order to address the challenge of how to model the temporal dependencies from the input data, we designed a Transformer-based multimodal encoder. This way, the attention mechanisms from the Transformer are used to model the temporal relations in the input modalities.

Our decoder was designed to address the second challenge, i.e. how to aggregate multimodal information. For this, our decoder uses a cross-attention mechanism that aggregates the information from the different modalities, using as input the representations given by the encoder. Using cross-attention implies that multimodal aggregation is done by performing a weighted sum of the representations of the different modalities. Those weights change dynamically depending on the input, and can be understood as the model identifying the most relevant modalities.

For the third challenge, taking into account past predictions, our decoder infers emotions in an auto-regressive manner. This means that to predict the current emotion, it uses as inputs the previous predictions.

We evaluated our approach on a state-of-the-art dataset, obtaining results that improved the baseline provided by the authors of the dataset, in terms of Root Mean Square Error (RMSE) and Concordance Correlation Coefficient (CCC). Moreover, we showed the validity of our decoder design, by replacing it with other architectures. Specifically, we replaced our Transformer-based decoder with a Fully-Connected Network (FCN) and a Long Short-Term Memory (LSTM) network, to process the representations given by the encoder. The experimental results showed that the performance of our solution was better than the alternative approaches, in terms of RMSE and CCC.

The work done for this contribution demonstrates that using attention mechanisms to aggregate multimodal information is useful, as this could

lead to improved performance. It also shows that it is important to take past predictions into account when doing time-continuous emotion recognition.

7.1.4 Accommodating Missing Modalities

Our last contribution is an approach to perform time-continuous multimodal emotion recognition robust to missing modalities. For this, we used the same architecture developed for our previous contribution, since an attention-based model is well-suited to handle missing modalities. In fact, we illustrated how this type of architecture can *naturally* accommodate missing modalities by attending to the remaining ones.

Through different experiments, we demonstrated that our model was indeed capable of handling missing modalities without any architectural change. Specifically, we showed that at test time, the model could still work even with a modality absent. However, we noted that there were modalities that when absent, significantly decreased the performance of the model, in terms of RMSE and CCC. From this, we concluded that there were modalities from which the model was extracting more information, thus they were *important* modalities, i.e. the most discriminant modalities.

To alleviate the performance decrease when an important modality is missing, we introduced an optimized training strategy, which consisted of hiding part of the time the important modalities during training. This way, we forced the model to use information from the *weaker* modalities (less discriminant modalities), such that when an important modality is missing, the model can still rely on the other ones.

We experimentally tested our ideas, finding that using the optimized training strategy led to improved performance in terms of CCC and RMSE when a modality was missing, compared to the performance when the same modality was missing and the model was trained using a standard approach. Moreover, given that with the optimized training strategy the model learns to use more information from the weak modalities, there was also an increase in performance when all the modalities were present.

The results show that our Transformer-based architecture is capable of accommodating missing modalities without the need of architectural changes, and it is partially robust to missing modalities even without any special training strategy. We also showed that robustness to missing modalities can be improved by hiding the important modalities during training.

Having a model that can accommodate missing modalities is important

because in real-life applications there can be cases when a modality may be missing. For example, a sensor could be faulty, or the user may decide to intentionally disconnect one of the inputs, say a video camera, for privacy reasons.

7.2 Limitations and Perspectives

We now discuss some limitations of our contributions, presenting some perspectives to overcome those limitations and to further expand our work.

7.2.1 Limitations on Datasets

As described in Section 2.4, different criteria were used to select the datasets that we employ to test our contributions. Some of these criteria included the type of sensors used to acquire the signals and the stimulus used to generate the emotions in the subjects from whom the different signals were acquired. The participants in the datasets selected under these criteria were healthy people, with ages between 18 and 40 years. However, as we envision the emotion recognition techniques developed in this thesis as part of a global health monitoring system for frail people, we consider a limitation to our work the fact that we did not test our model with samples taken from frail people, but tested only on healthy young people. Although we still consider our results valid towards our general goal of monitoring the emotional health of frail people, an interesting perspective is to study the differences that might exist between signals acquired from frail and younger healthy people. For example, physiological manifestations or facial expressions might be more pronounced at different age ranges.

Another limitation of the selected datasets is in the stimuli used to produce emotions in the subjects. We tried to have some variability regarding this, by using datasets that employed audio-visual stimulus for the first part of the thesis, and using a dataset that acquired signals from subjects under stress-induced situations for the second part of the thesis. Nevertheless, we are far from covering the whole range of emotion-inducing stimuli that a frail person might experience, like talking with a family member, receiving good or bad news, health-related issues, and others. Future work should take this into account, studying if models trained with data elicited with certain types of stimuli can function with data elicited with other types of stimuli.

Finally, our approaches were tested with a limited quantity of datasets:

two datasets when we worked only with physiological signals and one dataset when we worked with multimodal signals. Future work should include testing our models with more datasets to better evaluate the performance of our approaches.

7.2.2 General Models vs. Specific Models for Emotion Recognition

During this thesis, for each task we addressed, we designed a single architecture to recognize the two dimensions of emotions with which we worked: arousal and valence. And then, we trained the architecture separately for each dimension. We did this because we aimed to design a general architecture to predict emotions, rather than designing specific solutions for each emotion dimension. Then, during training, this architecture specializes in a particular emotion dimension.

However, another solution will be to train a single multi-tasking model, capable of predicting the different emotion dimensions at the same time. And going to the other extreme, another solution is to design specific architectures for each emotion dimension. The latter option is especially used in works that participate in challenges, where the objective is to obtain the best performance in terms of a defined metric [87, 143].

From this, a perspective that emerges is that we should study and compare the performances of the described approaches, understanding the strengths and weaknesses of each of them. This probably will help to answer some interesting questions, at least from a machine-learning point of view: Are the different emotion dimensions very independent, so using different architectures for each dimension works better? Or are these dimensions deeply related so a multi-task approach is more convenient?

Another aspect of generalization in contrast to specialization is having a general model that works for *all* users as opposed to a personalized specific model for each user. For the latter option, it means that the model is personalized by training it with data from the specific user, maybe in a few-shot learning fashion. In this thesis, we did not work on specialized models, because that implies that people have to label their own data, which we see as a barrier to deploying a system in real-world scenarios, especially for frail people. Nevertheless, as it is expected that a personalized model be more accurate than a general one, an interesting perspective to explore is to investigate if a personalized system is feasible, and if there are ways that such systems could be practical to deploy, searching for ways to easily label

data, for example.

7.2.3 Alternative Ways to Pre-Train the Models

When we worked only with physiological signals, we developed a pre-training strategy consisting of masking some parts of the signal and predicting those masked parts. In a way, this can be seen as a pre-training based on denoising: we added noise to the signal by putting zeros in some parts of the signal and then tried to reconstruct the original signal. In any case, our pre-training strategy falls under the category of a generative strategy [216]. But as we saw in Section 3.2.5.1, there are other strategies, notably predictive strategies, and contrastive learning. Predictive strategies typically consist of setting the pre-training task as a classification problem, for example by applying a transformation to the input and identifying which transformation took place. In fact, this is the strategy employed in the approach of Sarkar and Etemad [162], with which we compare our approach when we process single physiological signals in Chapter 3. Contrastive learning consists of generating representations from the data such that these representations are closer for related data (positive examples) and farther for unrelated data (negative examples).

We consider it an interesting perspective to explore those other types of pre-training strategies, especially contrastive learning, as it has demonstrated great success in other domains. The challenge with this strategy is to identify a way to build positive and negative samples. One way to do this is to take advantage of the sequential nature of physiological signals, and explore techniques similar to contrastive predicting coding [195]. In contrastive predicting coding the idea is to use the signal up to a specific time to predict representations that come after that time. This way, these generated representations should be closer to the real subsequent representations, and farther from any other representation of the signal. Other options include creating positive examples by transforming the original signal [39]. In any case, we believe contrastive learning, and other pre-training strategies, are an attractive avenue to investigate.

7.2.4 Using Characteristics of the Physiological Signals

A characteristic of our approach to processing physiological signals for emotion recognition is that we do not use parameters extracted from the signals, but we use raw signals instead. Nevertheless, it may be useful to use certain characteristics of the signals to improve the performance of

the models, identifying these characteristics using the knowledge about physiological signals that exist in medical and psychological domains. In a way, this could be seen as incorporating external knowledge into the system. For example, some psychological studies have investigated the influence of emotions on heart rate and heart rate variability [207]. Then, a way to incorporate this knowledge could be pre-training the model such that it predicts those quantities, a technique that has been tested for tasks in the medical domain [201]. We think that exploring this and other techniques to incorporate external (expert) knowledge in our model could be interesting.

In addition to this, we saw that our architecture gives more weight to certain elements of the Electrocardiogram (ECG) waveform (see Section 3.4.2.5 and Figure 3.11). It could be interesting to understand why those parts of the signal are more important, and if this has a relation with how emotions influence cardiac responses. Such understanding could be useful in tuning the model to make it better adapted for processing ECG with the aim of emotion recognition.

7.2.5 Characteristics of our Encoder-Decoder Architecture

To process multimodal signals to perform time-continuous emotion recognition, we developed a Transformer-based encoder-decoder architecture, which is described in Chapter 5. A limitation of our encoder-decoder approach is that having the encoder and the decoder increases the number of parameters of the model, thus it takes more time to train, requires more hardware resources, and may be more prone to overfitting than a smaller model. A way to reduce the number of parameters could be to use only the decoder. In this case, the decoder will not have a cross-attention module but will be formed only with self-attention layers. The input will be a sequence formed with the input signals and the target labels, in this way:

$$[x_1^1, \ldots, x_1^M, y_1, x_2^1, \ldots, x_2^M, y_2, \ldots, x_T^M, \ldots, x_T^M, y_T,], \tag{7.1}$$

where x_t^m is the feature of modality m at time-step t, and y_t is the label at that time-step. Then, for training, a strategy similar to next-token prediction can be used. In other words, the model can be trained to predict label y_t using as input the sequence up to that position.

These ideas are largely used in Language Models [28], and some authors have investigated its use with multimodal data [151]. We think it should be interesting to explore these approaches for multimodal emotion recognition, answering questions like if the quantity of data available in datasets with

labels of emotion is enough to train these models, if some type of pre-training will be necessary, and if so, how to design a pre-training task.

7.2.6 Unaligned Inputs for Multimodal Emotion Recognition

Our model developed to perform time-continuous emotion recognition was designed under the assumption that the multimodal inputs are aligned with each other. In other words, there is a fixed sampling rate at which all the features from the different modalities are extracted, so at any time t we have available the multimodal features corresponding to that time t. This is a limitation, as in the real world data may be extracted at different sample rates, there might be lagging in the data extraction and communication, and there might be other factors that produce unaligned multimodal data. Some authors have already addressed this issue, notably using cross-modal attention [190], but this approach is not scalable in terms of the number of modalities.

Future work could include extending our architecture to take alignment issues into account. For example, the cross-attention layers of our decoder can be adapted to attend to unaligned multimodal features given by the encoder, in a way emulating what is done in cross-modal attention [190], but with the advantage that it will have better scalability.

7.2.7 About Missing Modalities

To develop our training strategy to make our model more robust to missing modalities, described in Chapter 6, we did not take into account how probable each modality is to be missing. For example, it could be the case that important modalities have a low probability of being absent, or non-important modalities have a high probability of being missing. Moreover, we assumed that only one modality is missing at a time, whereas in real life this could not be the case. Thus, future work could study these different circumstances. It could be interesting to evaluate how our approach performs under these situations, and if a different training strategy is needed for these scenarios.

In addition, when we tested our approach with a missing modality, we did it by eliminating the modality from the input. However, it could be the case that a modality is not absent but is extremely noisy, or its values are spurious, and for this case, we assumed that there exists a perfect method

to identify these data as invalid, such that the model can ignore it. Although there are ways to identify if a modality falls under those cases (i.e. is not valid), for example by checking if its values are correlated with the values of the other modalities [136], these methods are not perfect. Therefore, future work remains in finding ways to incorporate these techniques into our approach, taking into account their accuracy, and how their possible errors would impact our approach. This way, our approach will become more robust in handling different scenarios that can arise regarding missing modalities, noisy inputs, and spurious values.

7.2.8 Giving Incorrect Recognition Results

If we imagine an emotion recognition system as part of a global system that monitors and offers services to frail people, then perfect accuracy is required from the recognition system. Detecting incorrect emotions might lead to incorrect behavior of the global system which can have dramatic consequences in these particular applications, like the global system suggesting to take a medicine that is not necessary. A way to alleviate this is to measure the uncertainty of the predictions made by the emotion recognition system. This way, the global system may accept only predictions that are considered correct with a high level of confidence. Note that using softmax confidence is not a good measurement of uncertainty [146], thus it is necessary to study how to adapt in our approaches methods to measure uncertainty such as Bayesian deep learning [75].

Even if only high-confidence predictions are used, the system can still make mistakes. Therefore, before deploying systems that monitor the emotions of frail people, it is important to understand the consequences for the user when the system does not work properly, and to identify, study, and find ways to mitigate these consequences before deploying such systems in the real world.

7.2.9 Ethical Implications

One of the ethical implications of our research is the risk of the use of our work in negative applications, like behavior manipulation. We acknowledge this potential risk, and recognize that the problem of using research for negative applications is in fact a problem that concerns many domains in Artificial Intelligence. Therefore, we advocate that the research community should seek that governmental entities legislate and control the use of Artificial Intelligence technologies, to avoid their use for harmful purposes.

Moreover, any monitoring application, especially if we are monitoring emotions, should be deployed with the full consent and understanding of the user. Therefore, efforts should be put into making people understand how the technology works, its potential risks, and the measures taken to mitigate those risks.

Another important issue is privacy considerations. In this regard, one of our contributions can be used to better ensure privacy: our approach that makes the model robust to missing modalities. To see how, we can imagine that a user may not want that a certain modality, that he or she considers invasive, to be captured and used by the system. In this case, the user can intentionally disable the modality considered as invasive and the system could keep working with the remaining modalities. Nevertheless, more work is needed to better ensure privacy, especially since some data sources used to recognize emotions may contain sensitive medical information. Using Machine Learning (ML) frameworks that are designed to be more privacy-oriented should be investigated. One of such frameworks that is well suited for smart environments is Federative Learning [132]. Within this framework, coordinated learning and inference could occur at each object of the smart environment, without having the gathered data leaving the object, thus better preserving privacy. For these reasons, an interesting perspective to investigate is how to adapt approaches like the ones developed in this thesis to work in a federative way.

Another important aspect that should be considered is energy use. Future work could include studying the energy consumption of our different approaches when training and when doing inference. With those findings, ways to make our architecture more energy-efficient could be investigated. For example, it may be deemed not necessary to use all the modalities when doing inference, as acceptable performance may be achieved with fewer modalities. Using fewer modalities implies less data to process and therefore, less energy consumption.

APPENDIX A

DATASETS

A.1 DREAMER Dataset

In [102], Katsigiannis and Ramzan introduce the DREAMER dataset, which consists in Electroencephalogram (EEG) and ECG signals recorded during emotional episodes provoked by audio-visual stimuli. Their goal is to use portable, wearable, wireless, low-cost, and off-the-shelf equipment to register the EEG and ECG signals, so the application of algorithms related to emotion and physiological signals can be expanded into everyday scenarios.

A.1.1 Acquisition Setup

Audio-visual stimuli were used to elicit emotional reactions from the participants. For this, 18 film clips were utilized, each clip containing scenes that have been shown to be capable of evoking a specific emotion. Specifically, each two of the 18 clips targeted one of the following emotions: amusement, excitement, happiness, calmness, anger, disgust, fear, sadness,

and surprise. The lengths of the clips were between 65 and 393 seconds, with an average of 199 seconds. 23 subjects participated in the experiments, aged between 22 and 33 years old. The total collected data amounts to a total of around 23 hours.

A.1.2 Provided Signals

The DREAMER dataset provides EEG and ECG signals obtained from commercial off-the-shelf devices. EEG signals were registered using an Emotiv EPOC system [1], providing the following 14 EEG channels: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4 (See Figure 2.8b). ECG signals were recorded using the Shimmer platform [29].

A.1.3 Labeling

After watching each clip, participants self-assessed the levels of arousal, valence, and dominance that they felt while watching the clip. To facilitate the annotations, Self-Assessment Manikins [138] were used. Arousal, valence and dominance were annotated on a scale of 1 to 5, ranging from uninterested/bored to excited/alert for arousal, unpleasant/stressed to happy/elated for valence, and helpless to empowered for dominance.

A.2 AMIGOS Dataset

The AMIGOS dataset was introduced by Miranda-Correa et al. in [135] with the aim of facilitating the multimodal study of the affective response of people. The dataset is built by recording different signals from participants while they are exposed to emotional fragments of movies.

A.2.1 Acquisition Setup

The 40 subjects that participated in the trials to build this dataset took part in two sets of experiments. In the first experiment, the 40 subjects individually watched 16 short videos with lengths between 51 to 150 seconds, with the average length of the videos being 86.7 seconds. In the second experiment, 37 of the 40 subjects watched 4 long videos with lengths between 14.1 to 23.58 minutes, with an average length of 20 minutes. The subjects watched the videos either alone or in groups.

1. https://www.emotiv.com/

The videos were selected so they produce emotional responses, with different videos covering a quadrant of the valence-arousal space, namely High Valence-High Arousal, High Valence-Low Arousal, Low Valence-High Arousal, and Low Valence-Low Arousal.

A.2.2 Provided Signals

The Amigos dataset provides the signals that we detail below.

Physiological Signals: In this dataset, three types of physiological signals are provided: ECG, EEG, and Electrodermal Activity (EDA). Similar to what was done for the DREAMER dataset, instead of using laboratory-grade instruments to acquire these signals, the authors preferred to use wearable low-cost off-the-shelf devices. In fact, they used the same equipment as the DREAMER authors, using the Emotive EPOC system to capture EEG signals, and the Shimmer platform to capture ECG signals. To capture EDA signals, an additional Shimmer board was used to extend the functionality of the Shimmer platform. The provided EEG channels are AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4 (See Figure 2.8b).

Video Recordings: The AMIGOS dataset provides frontal face recordings in HD quality. In addition to this, RGB and depth full-body videos were also captured.

A.2.3 Labeling

The authors of the AMIGOS dataset performed affective annotation using internal (self-assessment) and external annotations.

The self-assessment annotation was performed at the end of each trial. Participants selected discrete emotions and annotated the levels of valence, arousal, and dominance that they felt while watching each video. Also, they annotated if they liked or not the video, and how familiar they were with it, in a range from "Never seen it before" to "Know the video very well". Valence, arousal, and dominance were annotated on a scale of 1 to 9 using Self-Assessment Manikins [138], allowing participants to assess their arousal level from "very-calm" (1) to "very excited" (9), their valence level from "very negative" (1) to "very positive" (9), and their dominance level from "overwhelmed with emotions" (1) to "in full control of emotions" (9). Regarding discrete emotions, the subjects had to select at least one of the following: neutral, disgust, happiness, surprise, anger, fear, and sadness.

For the external annotations, the frontal face videos were used to annotate values of arousal and valence. Specifically, the videos were cropped so only a squared region around the face was visible for each participant. Then, the videos were split into 20-second clips. Three annotators rated for each clip, on a scale of -1 to 1, the perceived levels of arousal and valence.

A.3 Ulm-Trier Social Stress Test (ULM-TSST)

The Ulm-Trier Social Stress Test (ULM-TSST) dataset was introduced by Stappen et al. [179, 180] for the Multimodal Sentiment Analysis (MuSe) 2021 challenge, which was held as part of the ACM Multimedia 2021 conference. The same dataset was also used in the MuSe 2022 challenge [42]. Specifically, the dataset was used for the MuSe-Stress sub-challenge, which was a regression task of time-continuous values of arousal and valence.

A.3.1 Acquisition Setup

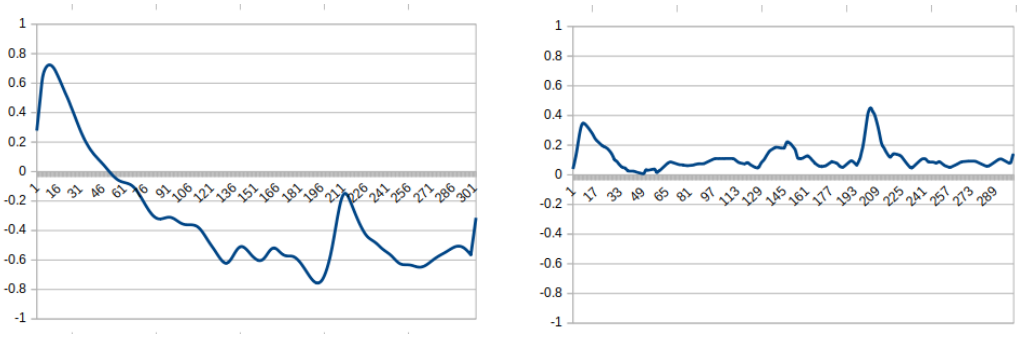
The data from the ULM-TSST dataset were captured from subjects during a stress-induced situation, following the Trier Social Stress Test (TSST) protocol [107]. TSST induces stress by simulating a job interview, where participants have to give a five-minute free speech oral presentation in front of two interviewers, who remain silent during the presentation. In total 69 subjects participated in the experiment, with ages between 18 and 39 years, generating around 6 hours of data.

A.3.2 Provided Signals and Features

The dataset provides audio and video recordings of each participant's five-minute speech, and it also provides a transcript of the speech. In addition, four physiological signals are recorded: EDA, ECG, Respiration (RESP), and Beats per Minute (BPM). From those physiological signals, only the last three are provided, since EDA is used to build the arousal ground truth as we shall see later.

Besides the raw signals, the authors of the ULM-TSST dataset provided the features described below. All the provided features are obtained at 0.5-second intervals, such that they are aligned to the labels.

eGeMAPS (audio): The authors used the openSMILE toolkit [65] to extract 88 extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)



(a) Arousal ground truth.

(b) Valence ground truth.

Figure A.1 – Example of arousal (Fig. A.1a) and valence (Fig. A.1b) ground-truth values of a sample from the ULM-TSST dataset.

features [64] (see Section 2.2.3.2 for more information on eGeMAPS).

DeepSpectrum (audio): DeepSpectrum features [8] are extracted using Convolutional Neural Network (CNN) pre-trained on images. For this, image representations of the sound are used (i.e. spectrograms). The obtained features are 1024-dimensional vectors.

VGGFace2 (video): These 512-dimensional features are extracted using a ResNet50 network [86] trained on the VGGFace 2 dataset [34] for the task of face recognition. The inputs to obtain these features are faces that are automatically extracted from the videos using the MTCNN model [217].

AU (video): As described in Section 2.2.3.1, it is possible to use Facial Action Coding System (FACS) [53] to deconstruct facial expressions into distinct muscle movements called Action Units (AUs). The authors of the ULM-TSST dataset used the Py-Feat toolbox [40] to automatically obtain 20 different AU, using as input faces extracted in the same way as it was done for the VGGFace2 features.

BERT (text): Text features are extracted using the BERT model [51], obtaining a 768-dimensional vector.

Physiological Signals: Physiological signals are downsampled to 2Hz and smoothed with a Savitzky-Golay filter. Then, the three signals (ECG, RESP, and BPM) are concatenated to form a 3-value feature vector.

A.3.3 Labeling

The ULM-TSST dataset was annotated by three raters for the emotional dimensions of arousal and valence. The annotation for both dimensions is done every 0.5 seconds, using values in the -1 to 1 range.

The ground truth for valence was obtained by aggregating the scores of the three annotators using the Rater Aligned Annotation Weighting (RAAW) method from the MuSe-Toolbox [181]. RAAW is a method to aggregate annotations from various raters, addressing the rater lag that might be present, and performing weighted aggregation by assigning weights to the annotations of each rater according to their agreement with the mean of all others. Figure A.1b shows an example of the obtained valence ground truth of a sample.

The ground truth for arousal is obtained in a similar way, with the difference that the annotations with the lower inter-rater agreement are discarded and replaced with the EDA signal from the corresponding subject [16]. The authors do this because EDA signals are known to be a good indicator of arousal [36]. Figure A.1a shows an example of the obtained arousal ground truth of a sample.

Of the 69 samples, the authors only provide labels for the 55 samples that were selected by them as train and validation sets. The labels for the other 14 samples, which form the test set, are held by the authors for the purpose of the challenge.

APPENDIX B

SETTING THE DATASETS FOR SELF-SUPERVISED PRE-TRAINING

B.1 Pre-Training Setup

Note that AMIGOS and DREAMER datasets are part of the pre-training datasets and the evaluation datasets for ECG signals, and the AMIGOS dataset is part of the pre-training datasets and the evaluation datasets for EEG signals. Therefore, it is necessary to be careful to avoid using the same samples to pre-train and test the model.

In our experiments, we perform a pre-training process for each evaluation scenario when working with ECG signals. On the other hand, when working with EEG signals, we use the same pre-trained model for all the evaluation datasets, with a caveat when using AMIGOS, as we shall see. This means that we do the following:

— Pre-training with ECG signals, for AMIGOS as evaluation dataset.

— Pre-training with ECG signals, for DREAMER as evaluation dataset.

— Pre-training with EEG signals, for AMIGOS and DREAMER as evaluation dataset.

In order to avoid testing with a sample that was already used for pre-training, we perform the procedure described below. This procedure does not need to be done when DREAMER with EEG signals is used as evaluation set, since DREAMER is not used at all for pre-training EEG signals. First, each evaluation dataset is divided into two halves. Then, we pre-train two models: one model is pre-trained using all the other pre-training datasets plus the first half of the evaluation dataset. The second model is also pre-trained with all the other datasets but now uses the other half of the evaluation dataset. At evaluation time, if we are testing a certain sample from the evaluation dataset, we make sure to use the model that was pre-trained with the half of the evaluation dataset that does not contain that sample. Appendix B.2 explains how each half of the evaluation datasets are created.

An alternative to pre-training multiple models is to pre-train a single model for ECG signals and a single model for EEG signals. These models could be pre-trained without using any of the evaluation datasets, thus avoiding the possibility of using the same samples to pre-train and test the model. Then the pre-trained ECG and EEG models would be fine-tuned independently for each evaluation dataset. In this case, no part of the AMIGOS or DREAMER dataset would be used for pre-training. This approach was not adopted in order to use as much data as possible, because with our strategy, we can incorporate data that is not part of the *current* evaluation set in the pre-training datasets. For example, with AMIGOS as evaluation dataset with ECG signals, it is possible to use half of the AMIGOS dataset and the whole DREAMER dataset, with the rest of the pre-training datasets, to pre-train the model. It could be interesting to explore how much pre-training data is necessary for the model to perform well. This could help determine if it is necessary to employ our strategy or if a simpler approach of not using the evaluation datasets for pre-training is enough. We leave these questions as future work.

Table B.1 shows the number of 10-second segments that are used to pre-train the model. Note that these quantities are the ones used for each of the two pre-trained models for each evaluation scenario.

Evaluation Scenario		
Signal	Evaluation Dataset	Segments
ECG	AMIGOS	83,401
ECG	DREAMER	98,295
EEG	AMIGOS, DREAMER	45,805

Table B.1 – Number of 10-second segments used for each pre-trained model on each evaluation scenario.

B.2 Fine-Tuning Setup

To avoid using the same samples for pre-training and evaluation, when fine-tuning and evaluating the model we follow the strategy described below.

To evaluate our approach, we use 10-fold cross-validation, so each evaluation dataset is divided into 10 folds. Recall that for each evaluation scenario, except for DREAMER with EEG signals, we pre-train two models. Specifically, for each evaluation dataset D and for each type of physiological signal S, we pre-train two signal encoders, noted as $SE_1^{D,S}$ and $SE_2^{D,S}$. To pre-train $SE_1^{D,S}$, we use the folds 1 to 5 of the evaluation dataset D and the rest of the pre-training datasets, using the physiological signal S. Correspondingly, the second signal encoder $SE_2^{D,S}$ is pre-trained with the folds 6 to 10 of the evaluation dataset D and the rest of the pre-training datasets, using the signal S. Then, we fine-tune a model initialized with the weights from $SE_2^{D,S}$, if the model is tested on folds 1 to 5 for emotion recognition on the evaluation dataset D using the signal S. Likewise, to test a model on folds 6 to 10 for emotion recognition on the evaluation dataset D using the signal S, the model is initialized with the weights from $SE_1^{D,S}$. This method allows us to pre-train, fine-tune and test the model in a more efficient way than pre-training 10 different models, one for each fold, while retaining complete separations between training and testing data. When we fine-tune DREAMER with EEG signals, we choose the first pre-trained EEG signal encoder, i.e. we use the model $SE_1^{\text{AMIGOS, EEG}}$.

BIBLIOGRAPHY

[1] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, July 2015. 36

[2] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors*, 21(4):1249, January 2021. 29, 30

[3] Gustavo Aguilar, Viktor Rozgic, Weiran Wang, and Chao Wang. Multimodal and Multi-view Models for Emotion Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 991–1002, Florence, Italy, July 2019. Association for Computational Linguistics. 144, 145, 146

[4] David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. Attention Networks for Multi-Task Signal Analysis. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 184–187, July 2020. 49

[5] Mahmoud I. Al-kadi, M. B. I. Reaz, and M. A. Mohd Ali. Compatibility of mother wavelet functions with the electroencephalographic signal. In *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*, pages 113–117, December 2012. 50

[6] Salma Alhagry, Aly Aly, and Reda A. Emotion Recognition based on EEG using LSTM Recurrent Neural Network. *International Journal of Advanced Computer Science and Applications*, 8(10), 2017. 30, 47

[7] Murray Alpert and Anna Rosen. A semantic analysis of the various ways that the terms "affect," "emotion," and "mood" are used. *Journal of Communication Disorders*, 23(4):237–246, August 1990. 2

[8] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. Snore Sound Classification Using Image-Based Deep Spectrum Features. In *Interspeech 2017*, pages 3512–3516. ISCA, August 2017. 183

[9] Arjun Arjun, Aniket Singh Rajpoot, and Mahesh Raveendranatha Panicker. Introducing Attention Mechanism for EEG Signals: Emotion Recognition with Vision Transformers. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5723–5726, November 2021. 49, 50

[10] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, November 2010. 84, 85

[11] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv:1607.06450 [cs, stat]*, July 2016. 34, 62, 68

[12] Dominik R. Bach, Samuel Gerster, Athina Tzovara, and Giuseppe Castegnetti. PsPM-RRM1-2: SCR, ECG, respiration and eye tracker measurements in response to electric stimulation or visual targets, September 2019. 66

[13] Jo-Anne Bachorowski. Vocal Expression and Perception of Emotion. *Current Directions in Psychological Science*, 8(2):53–57, April 1999. 19

[14] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 49, 62

[15] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv:1803.01271*, April 2018. 121

[16] Alice Baird, Lukas Stappen, Lukas Christ, Lea Schumann, Eva-Maria Messner, and Björn W. Schuller. A Physiologically-Adapted Gold Standard for Arousal during Stress. In *Proceedings of the 2nd on*

Multimodal Sentiment Analysis Challenge, pages 69–73, New York, NY, USA, October 2021. Association for Computing Machinery. 184

[17] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, February 2019. 84

[18] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv:2003.05991*, April 2021. xx, 55, 91

[19] Philip Bard. A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system. *American Journal of Physiology-Legacy Content*, 84(3):490–515, April 1928. 10

[20] Lisa Feldman Barrett. The Future of Psychology: Connecting Mind to Brain. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 4(4):326–339, July 2009. 11

[21] Yakoub Bazi, Laila Bashmal, Mohamad Al Rahhal, Reham Dayil, and Naif Ajlan. Vision Transformers for Remote Sensing Image Classification. *Remote Sensing*, 13:516, February 2021. xix, 36

[22] Behnam Behinaein, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. A Transformer Architecture for Stress Detection from ECG. In *2021 International Symposium on Wearable Computers*, pages 132–134, Virtual USA, September 2021. ACM. 49

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. 52

[24] Uttaran Bhattacharya, Christian Roncal, Trisha Mittal, Rohan Chandra, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Take an Emotion Walk: Perceiving Emotions from Gaits Using Hierarchical Attention Pooling and Affective Mapping. In *Computer Vision – ECCV 2020*, pages 145–163, Cham, 2020. Springer International Publishing. 18

[25] Frans A Boiten. The effects of emotional behaviour on components of the respiratory cycle. *Biological Psychology*, 49(1):29–51, September 1998. 23

[26] Wolfram Boucsein. *Electrodermal Activity*. Springer US, Boston, MA, 2012. xix, 24, 25

[27] Perry S. Braun and Bobbi Kloss. Wellness Programs–Emotional Wellness. *Employee Benefit Plan Review*, 72(1):27–28, September 2017. 1

[28] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv:2005.14165*, July 2020. 175

[29] Adrian Burns, Barry R. Greene, Michael J. McGrath, Terrance J. O'Shea, Benjamin Kuris, Steven M. Ayer, Florin Stroiescu, and Victor Cionca. SHIMMER™ – A Wireless Sensor Platform for Noninvasive Biomedical Research. *IEEE Sensors Journal*, 10(9):1527–1534, September 2010. 180

[30] Cong Cai, Yu He, Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao, Mingyu Xu, and Kexin Wang. Multimodal Sentiment Analysis based on Recurrent Neural Network and Multimodal Attention. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 61–67, Virtual Event China, October 2021. ACM. 49

[31] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep Adversarial Learning for Multi-Modality Missing Data Completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1158–1166, New York, NY, USA, July 2018. Association for Computing Machinery. xxii, 146, 148, 149

[32] Erik Cambria, Andrew Livingstone, and Amir Hussain. The Hourglass of Emotions. In *Cognitive Behavioural Systems*, volume 7403, pages 144–157. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. xix, 12, 13

[33] W. B. Cannon. The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, 39:106–124, 1927. 10

[34] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE Computer Society, May 2018. 183

[35] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A Short Note on the Kinetics-700 Human Action Dataset. *arXiv:1907.06987*, October 2022. 42

[36] Delphine Caruelle, Anders Gustafsson, Poja Shams, and Line Lervik-Olsen. The use of electrodermal activity (EDA) measurement to understand consumer emotions – A literature review and a call for action. *Journal of Business Research*, 104:146–160, November 2019. 23, 184

[37] Debatri Chatterjee, Rahul Gavas, and Sanjoy Kumar Saha. Exploring Skin Conductance Features for Cross-Subject Emotion Recognition. In *2022 IEEE Region 10 Symposium (TENSYMP)*, pages 1–6, July 2022. 25

[38] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer Encoder With Multi-Modal Multi-Head Attention for Continuous Affect Recognition. *IEEE Transactions on Multimedia*, 23:4171–4183, 2021. 60, 112, 113, 114, 121, 123, 131

[39] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, November 2020. 174

[40] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J. Chang. Py-Feat: Python Facial Expression Analysis Toolbox, March 2023. 183

[41] Woan-Shiuan Chien, Huang-Cheng Chou, and Chi-Cun Lee. Self-assessed Emotion Classification from Acoustic and Physiological Features within Small-group Conversation. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 230–239, Montreal QC Canada, October 2021. ACM. 49

[42] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller. The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 5–14, New York, NY, USA, October 2022. Association for Computing Machinery. 129, 131, 133, 134, 182

[43] Paul Compagnon, Grégoire Lefebvre, Stefan Duffner, and Christophe Garcia. Learning personalized ADL recognition models from few raw data. *Artificial Intelligence in Medicine*, 107:101916, July 2020. 3

[44] Géraldine Coppin and David Sander. Chapter 1 - Theoretical approaches to emotion and its measurement. In *Emotion Measurement (Second Edition)*, pages 3–37. Woodhead Publishing, January 2021. 5, 14, 18, 20

[45] James Crowley and Joëlle Coutaz. An Ecological View of Smart Home Technologies. In *European Conference on Ambient Intelligence*, pages 1–16. Springer,, November 2015. 3

[46] Julien Cumin and Grégoire Lefebvre. A Priori Data and A Posteriori Decision Fusions for Human Action Recognition. In *11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, Roma, Italy, March 2016. 85

[47] Suresh Dara and Priyanka Tumma. Feature Extraction By Using Deep Learning: A Survey. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1795–1801, March 2018. 46

[48] Charles Darwin. *The Expression of the Emotions in Man and Animals*. John Murray., 1872. 18

[49] Federico Del Pup and Manfredo Atzori. Applications of Self-Supervised Learning to Biomedical Signals: Where are we now. *TechRxiv*, April 2023. 54, 55

[50] Sylvain Delplanque, Didier Grandjean, Christelle Chrea, Géraldine Coppin, Laurence Aymard, Isabelle Cayeux, Christian Margot, Maria Velazco, David Sander, and Klaus Scherer. Sequential Unfolding of Novelty and Pleasantness Appraisals of Odors: Evidence From Facial Electromyography and Autonomic Reactions. *Emotion (Washington, D.C.)*, 9:316–28, July 2009. 21, 22

[51] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 30, 61, 62, 183

[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]*, October 2020. 49, 50

[53] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978. 18, 183

[54] P. Ekman, E. R. Sorenson, and W. V. Friesen. Pan-cultural elements in facial displays of emotion. *Science (New York, N.Y.)*, 164(3875):86–88, April 1969. 12

[55] Paul Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, 19:207–283, 1971. 11

[56] Paul Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384–392, April 1993. 18

[57] Paul Ekman. Basic Emotions. In *Handbook of Cognition and Emotion*, chapter 3, pages 45–60. John Wiley & Sons, Ltd, 1999. 11, 12, 13

[58] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, February 1971. 11

[59] Rayan Elalamy, Marios Fanourakis, and Guillaume Chanel. Multimodal emotion recognition using recurrence plots and transfer learning on physiological signals. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7, September 2021. 104

[60] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why Does Unsupervised Pre-training Help Deep Learning? In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, March 2010. 52

[61] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, May 2022. 53

[62] Eylül Ertay, Hao Huang, Zhanna Sarsenbayeva, and Tilman Dingler. Challenges of Emotion Detection Using Facial Expressions and Emotion Visualisation in Remote Communication. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pages 230–236, New York, NY, USA, September 2021. Association for Computing Machinery. 5

[63] Eurostat. Ageing Europe - statistics on housing and living conditions. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Ageing_Europe_-_statistics_on_housing_and_living_conditions. 3

[64] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, April 2016. 19, 130, 158, 183

[65] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462, New York, NY, USA, October 2010. Association for Computing Machinery. 182

[66] Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1437–1446. PMLR, July 2018. 68

[67] Beverley Fehr and James A. Russell. Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, 113(3):464–486, September 1984. 10

[68] Panagiotis Paraskevas Filntisis, Niki Efthymiou, Gerasimos Potamianos, and Petros Maragos. Emotion Understanding in Videos Through Body, Context, and Visual-Semantic Embedding Loss. In *Computer Vision – ECCV 2020 Workshops*, pages 747–755, Cham, 2020. Springer International Publishing. 25

[69] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936. 45

[70] W. Friesen and P. Ekman. EMFACS-7: Emotional Facial Action Coding System, 1983. 18

[71] Sean A. Fulop and Kelly Fitz. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *The Journal of the Acoustical Society of America*, 119(1):360–371, January 2006. 89

[72] Valentin Gabeur, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. Masking Modalities for Cross-modal Video Retrieval. In *2022 IEEE/CVF Winter Conference on Applications of Computer*

Vision (WACV), pages 2111–2120, Waikoloa, HI, USA, January 2022. IEEE. 151

[73] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal Transformer for Video Retrieval. In *Computer Vision – ECCV 2020*, volume 12349, pages 214–229, Cham, 2020. Springer International Publishing. 116, 118, 121, 122, 151

[74] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs Late Fusion in Multimodal Convolutional Neural Networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6, July 2020. 85

[75] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, pages 1050–1059, New York, NY, USA, June 2016. JMLR.org. 177

[76] Deepak Ghimire and Joonwhoan Lee. Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines. *Sensors*, 13(6):7714–7734, June 2013. 29

[77] Martin Gjoreski, Blagoj Mitrevski, Mitja Luštrek, and Matjaž Gams. An Inter-domain Study for Arousal Recognition from Physiological Signals. *Informatica*, 42(1), March 2018. 22, 43

[78] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–220, June 2000. 66

[79] Lucas Goncalves and Carlos Busso. AuxFormer: Robust Approach to Audiovisual Emotion Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7357–7361, May 2022. 150

[80] Ping Gong, Heather T. Ma, and Yutong Wang. Emotion recognition based on the multiple physiological signals. In *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 140–143, June 2016. 86

[81] Ayoub Hajlaoui, Mohamed Chetouani, and Slim Essid. EEG-based Inter-Subject Correlation Schemes in a Stimuli-Shared Framework: Interplay with Valence and Arousal. *arXiv:1809.08273*, September 2018. 5, 29

[82] Ayoub Hajlaoui, Mohamed Chetouani, and Slim Essid. Multi-task Feature Learning for EEG-based Emotion Recognition Using Group Nonnegative Matrix Factorization. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 91–95, September 2018. 25

[83] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, January 2023. 30

[84] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L. Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 International Symposium on Wearable Computers*, pages 45–49, New York, NY, USA, September 2020. Association for Computing Machinery. 57

[85] R. Harper and J. Southern. A Bayesian Deep Learning Framework for End-To-End Prediction of Emotion from Heartbeat. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. 47, 75

[86] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. 183

[87] Yu He, Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, and Yuan Cheng. Multimodal Temporal Attention in Sentiment Analysis. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 61–66, New York, NY, USA, October 2022. Association for Computing Machinery. 112, 131, 135, 173

[88] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural computation*, 9:1735–80, December 1997. 47

[89] Anne Horvers, Natasha Tombeng, Tibor Bosse, Ard W. Lazonder, and Inge Molenaar. Detecting Emotions through Electrodermal Activity in Learning Contexts: A Systematic Review. *Sensors (Basel, Switzerland)*, 21(23):7869, November 2021. 23, 24

[90] Y. Hsu, J. Wang, W. Chiang, and C. Hung. Automatic ECG-Based Emotion Recognition in Music Listening. *IEEE Transactions on Affective Computing*, 11(1):85–99, January 2020. 22, 29, 43, 44

[91] Jingzhao Hu, Chen Wang, Qiaomei Jia, Qirong Bu, Richard Sutcliffe, and Jun Feng. ScalingNet: Extracting features from raw EEG data

for emotion recognition. *Neurocomputing*, 463:177–184, November 2021. 47, 48, 59

[92] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, December 2006. 88, 89

[93] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 57–64, New York, NY, USA, October 2018. Association for Computing Machinery. 112, 131

[94] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Zhengqi Wen, Minghao Yang, and Jiangyan Yi. Continuous Multimodal Emotion Prediction Based on Long Short Term Memory Recurrent Neural Network. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 11–18, New York, NY, USA, October 2017. Association for Computing Machinery. 112

[95] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal Transformer Fusion for Continuous Emotion Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511, May 2020. 30, 49, 119

[96] Carroll E. Izard. *Human Emotions*. Plenum Press, New York, 1977. 12

[97] William James. What is an Emotion? *Mind*, 9(34):188–205, 1884. 10

[98] Haoning Kan, Jiale Yu, Jiajin Huang, Zihe Liu, and Haiyan Zhou. Self-supervised Group Meiosis Contrastive Learning for EEG-Based Emotion Recognition, August 2022. 55, 57

[99] Ashish Kapoor and Rosalind W. Picard. Multimodal affect recognition in learning environments. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 677–682, Hilton Singapore, November 2005. ACM. 144

[100] Arvid Kappas, Ursula Hess, and Klaus R. Scherer. Voice and emotion. In *Fundamentals of Nonverbal Behavior*, pages 200–238. Cambridge University Press, New York, NY, US, 1991. 19

[101] Kuldeep Singh Kaswan, Jagjit Singh Dhatterwal, Sanjay Kumar, and Amit Pandey. Chapter 4 - Industry 4.0 multiagent system-based

knowledge representation through blockchain. In *Artificial Intelligence and Industry 4.0*, pages 93–115. Academic Press, January 2022. 3

[102] Stamos Katsigiannis and Naeem Ramzan. DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, January 2018. 7, 8, 26, 36, 40, 42, 65, 75, 82, 86, 89, 98, 104, 179

[103] Murat Kaya, Mustafa Kemal Binli, Erkan Ozbay, Hilmi Yanar, and Yuriy Mishchenko. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. *Scientific Data*, 5(1):180211, December 2018. 66

[104] Pritam Khan, Priyesh Ranjan, and Sudhir Kumar. AT2GRU: A Human Emotion Recognition Model With Mitigated Device Heterogeneity. *IEEE Transactions on Affective Computing*, 14(2):1520–1532, April 2023. 104

[105] Aparna Khare, Srinivas Parthasarathy, and Shiva Sundaram. Self-Supervised Learning with Cross-Modal Transformers for Emotion Recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 381–388, January 2021. 90

[106] Eesung Kim and Jong Won Shin. DNN-based Emotion Recognition Based on Bottleneck Acoustic Features and Lexical Features. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6720–6724, May 2019. 20

[107] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer. The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993. 130, 182

[108] Paul R. Kleinginna and Anne M. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, 5(4):345–379, December 1981. 10

[109] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, January 2012. 36

[110] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel. Emotion Recognition and Its Applications. In *Human-Computer*

Systems Interaction: Backgrounds and Applications 3, pages 51–62. Springer International Publishing, Cham, 2014. 16

[111] Dimitrios Kollias and Stefanos Zafeiriou. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *arXiv:1910.04855 [cs, eess]*, September 2019. 37

[112] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040, March 2021. 37

[113] Beibei Kuang, Shenli Peng, Xiaochun Xie, and Ping Hu. Universality vs. Cultural Specificity in the Relations Among Emotional Contagion, Emotion Regulation, and Mood State: An Emotion Process Perspective. *Frontiers in Psychology*, 10, 2019. 11, 29

[114] Nandini Kumari, Shamama Anwar, and Vandana Bhattacharjee. Time series-dependent feature of EEG signals for improved visually evoked emotion classification using EmotionCapsNet. *Neural Computing and Applications*, February 2022. 75

[115] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 125–128. ISCA, September 2003. 29

[116] C. G. Lange. The mechanism of the emotions (Translated: Benjamin Rand). In *The Classical Psychologists (Pp. 672–684)*. Houghton Mifflin, 1885. 10

[117] Randy J. Larsen and Barbara L. Fredrickson. Measurement issues in emotion research. In *Well-Being: The Foundations of Hedonic Psychology*, pages 40–60. Russell Sage Foundation, New York, NY, US, 1999. 18

[118] Siddique Latif, Aun Zaidi, Heriberto Cuayahuitl, Fahad Shamshad, Moazzam Shoukat, and Junaid Qadir. Transformers in Speech Processing: A Survey, March 2023. 30

[119] Michael Lewis, Kiyoko Takai- Kawakami, Kiyobumi Kawakami, and Margaret Wolan Sullivan. Cultural Differences in Emotional Responses to Success and Failure. *International journal of behavioral development*, 34(1):53–61, January 2010. 5, 29

[120] Chao Li, Zhongtian Bao, Linhao Li, and Ziping Zhao. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. *Information Processing & Management*, 57(3):102185, May 2020. 104

[121] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 49

[122] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv:1807.05118 [cs, stat]*, July 2018. 68, 99, 130

[123] L. I. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, March 1989. 128

[124] Kristen A. Lindquist, Tor D. Wager, Hedy Kober, Eliza Bliss-Moreau, and Lisa Feldman Barrett. The brain basis of emotion: A meta-analytic review. *The Behavioral and brain sciences*, 35(3):121–143, June 2012. 11

[125] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris Metaxas. Multispectral Deep Neural Networks for Pedestrian Detection. In *Procedings of the British Machine Vision Conference 2016*, pages 73.1–73.13, York, UK, 2016. British Machine Vision Association. 85

[126] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, June 2022. 104

[127] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Emotion Recognition Using Multimodal Deep Learning. In *Neural Information Processing*, pages 521–529, Cham, 2016. Springer International Publishing. 91, 92

[128] Yiping Liu, Wei Sun, Xing Zhang, and Yebao Qin. Improving Dimensional Emotion Recognition via Feature-wise Fusion. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 55–60, New York, NY, USA, October 2022. Association for Computing Machinery. 112, 131, 135

[129] Matthew D. Luciw, Ewa Jarocka, and Benoni B. Edin. Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction. *Scientific Data*, 1(1):140047, November 2014. 66

[130] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are Multimodal Transformers Robust to Missing Modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022. 151

[131] M. Maithri, U. Raghavendra, Anjan Gudigar, Jyothi Samanth, Prabal Datta Barua, Murugappan Murugappan, Yashas Chakole, and U. Rajendra Acharya. Automated emotion recognition: Current trends and future perspectives. *Computer Methods and Programs in Biomedicine*, 215:106646, March 2022. 29

[132] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, April 2017. 178

[133] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology*, 14(4):261–292, December 1996. 14

[134] Wendy Mendes. Assessing autonomic nervous system activity. *Methods in Social Neuroscience*, pages 118–147, January 2009. 24

[135] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, April 2021. 7, 8, 26, 36, 40, 65, 67, 75, 86, 89, 98, 104, 180

[136] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1359–1367, April 2020. 143, 146, 147, 177

[137] Saif M. Mohammad. Chapter 11 - Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In *Emotion Measurement (Second Edition)*, pages 323–379. Woodhead Publishing, January 2021. 20

[138] Jon D. Morris. Observations: SAM: The self-assessment manikin: An efficient cross-cultural measurement of emotional response. *Journal of Advertising Research*, 35(6):63–68, 1995. 180, 181

[139] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention Bottlenecks for Multimodal Fusion. In *Advances in Neural Information Processing Systems*, volume 34, pages 14200–14213. Curran Associates, Inc., 2021. 85, 116, 117, 118, 120, 126

[140] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. ModDrop: Adaptive Multi-Modal Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, August 2016. 149

[141] Kengo Noguchi, Yuri Masaoka, Kanako Satoh, Nobumasa Katoh, and Ikuo Homma. Effect of Music on Emotions and Respiration. *The Showa University Journal of Medical Sciences*, 24(1):69–75, 2012. 23

[142] Robert Oostenveld and Peter Praamstra. The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, page 7, 2001. 22

[143] Ho-min Park, Ilho Yun, Ajit Kumar, Ankit Kumar Singh, Bong Jun Choi, Dhananjay Singh, and Wesley De Neve. Towards Multimodal Prediction of Time-continuous Emotion using Pose Feature Engineering and a Transformer Encoder. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 47–54, New York, NY, USA, October 2022. Association for Computing Machinery. 112, 131, 135, 173

[144] Srinivas Parthasarathy and Shiva Sundaram. Training Strategies to Handle Missing Modalities for Audio-Visual Expression Recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404, New York, NY, USA, December 2021. Association for Computing Machinery. 150

[145] Philipp C. Paulus, Giuseppe Castegnetti, and Dominik R. Bach. PsPM-HRM5: SCR, ECG and respiration measurements in response to positive/negative IAPS pictures, and neutral/aversive sounds, June 2020. 66

[146] Tim Pearce, Alexandra Brintrup, and Jun Zhu. Understanding Softmax Confidence and Uncertainty. *arXiv:2106.04972*, June 2021. 177

[147] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6892–6899, July 2019. 144, 145, 146

[148] Robert Plutchik. A General Psycheoevolutionary Theory of Emotion. In *Theories of Emotion*, pages 3–33. Academic Press, January 1980. xix, 12, 13

[149] E. Pranav, Suraj Kamal, C. Satheesh Chandran, and M.H. Supriya. Facial Emotion Recognition Using Deep Convolutional Neural Network. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 317–320, March 2020. 30

[150] Fano Ramparany, Iago F Trentin, Julien Cumin, and Olivier Boissier. Collaborative Homes. In *All the Agents Challenge (ATAC 2021)*, pages 7–12, October 2021. 4, 16

[151] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A Generalist Agent. *arXiv:2205.06175*, May 2022. 175

[152] Ajgou Riadh, Salim Sbaa, Ali Chamsa, and Abdelmalik taleb-ahmed. Novel Detection Algorithm of Speech Activity and the impact of Speech Codecs on Remote Speaker Recognition System. *WSEAS Transactions on Signal Processing*, Volume 10, 2014:309–319, July 2014. xix, 19

[153] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, April 2013. 37

[154] Kyle Ross, Paul Hungler, and Ali Etemad. Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data. *Journal of Ambient Intelligence and Humanized Computing*, October 2021. 55, 57, 67, 75, 91, 92, 104

[155] James Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39:1161–1178, December 1980. 14

[156] James A. Russell. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, 115(1):102–141, January 1994. 18, 19

[157] James A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, January 2003. xix, 14

[158] James A. Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805–819, May 1999. 27

[159] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, September 1977. 14

[160] Nastaran Saffaryazdi, Yenushka Goonesekera, Nafiseh Saffaryazdi, Nebiyou Daniel Hailemariam, Ebasa Girma Temesgen, Suranga Nanayakkara, Elizabeth Broadbent, and Mark Billinghurst. Emotion Recognition in Conversations Using Brain and Physiological Signals. In *27th International Conference on Intelligent User Interfaces*, pages 229–242, New York, NY, USA, March 2022. Association for Computing Machinery. 41

[161] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar. Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS). *IEEE Access*, 7:57–67, 2019. 22, 30, 40, 41, 47, 51, 59, 75, 86

[162] P. Sarkar and A. Etemad. Self-supervised ECG Representation Learning for Emotion Recognition. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. xx, xxiii, 40, 41, 55, 56, 57, 59, 67, 75, 76, 174

[163] P. Sarkar and A. Etemad. Self-Supervised Learning for ECG-Based Emotion Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3217–3221, May 2020. xx, 56

[164] Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE transactions on biomedical engineering*, 51(6):1034–1043, June 2004. 66

[165] Klaus R. Scherer, Tom Johnstone, and Gundrun Klasmeyer. Vocal expression of emotion. In *Handbook of Affective Sciences*, pages 433–456. Oxford University Press, New York, NY, US, 2003. 19

[166] Klaus R. Scherer and Agnes Moors. The Emotion Process: Event Appraisal and Component Differentiation. *Annual Review of Psychology*, 70(1):719–745, 2019. 11

[167] Jilt Sebastian and Piero Pierucci. Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts. In *Interspeech 2019*, pages 51–55. ISCA, September 2019. 20

[168] Jia-Lin Shen, Jeih-Weih Hung, and Lin-Shan Lee. Robust entropy-based endpoint detection for speech recognition in noisy environments. In *5th International Conference on Spoken Language Processing (ICSLP 1998)*, pages paper 0232–0. ISCA, November 1998. 41

[169] Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive Learning of Subject-Invariant EEG Representations for Cross-Subject Emotion Recognition. *IEEE Transactions on Affective Computing*, pages 1–1, 2022. 55

[170] Shimmer. ECG Respiration User Guide, 2018. xix, 25

[171] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A Review of Emotion Recognition Using Physiological Signals. *Sensors*, 18(7):2074, July 2018. 41

[172] Lin Shu, Yang Yu, Wenzhuo Chen, Haoqiang Hua, Qin Li, Jianxiu Jin, and Xiangmin Xu. Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet. *Sensors*, 20(3):718, January 2020. 43

[173] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig. Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity. *IEEE Transactions on Affective Computing*, pages 1–1, 2019. 41, 45

[174] Jonathan Sicsic, Bastian Ravesteijn, and Thomas Rapp. Are frail elderly people in Europe high-need subjects? First evidence from the SPRINTT data. *Health Policy*, 124(8):865–872, August 2020. 3

[175] S. Siddharth, T. Jung, and T. J. Sejnowski. Utilizing Deep Learning Towards Multi-modal Bio-sensing and Vision-based Affective Computing. *IEEE Transactions on Affective Computing*, pages 1–1, 2019. 75, 86, 88, 89, 90, 104

[176] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, April 2015. 88

[177] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, January 2012. 36

[178] T. Song, W. Zheng, P. Song, and Z. Cui. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, July 2020. 75

[179] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 5–14, Virtual Event China, October 2021. ACM. 37, 129, 182

[180] Lukas Stappen, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. MuSe 2021 Challenge: Multimodal Emotion, Sentiment, Physiological-Emotion, and Stress Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5706–5707, New York, NY, USA, October 2021. Association for Computing Machinery. 8, 37, 129, 182

[181] Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigell, Erik Cambria, and Björn W. Schuller. MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 75–82, New York, NY, USA, October 2021. Association for Computing Machinery. 184

[182] Benjamin Stephens-Fripp, Fazel Naghdy, David Stirling, and Golshah Naghdy. Automatic Affect Perception Based on Body Gait and Posture: A Survey. *International Journal of Social Robotics*, 9(5):617–641, November 2017. 18

[183] Ian Stevenson and Herbert S. Ripley. Variations in respiration and in respiratory symptoms during changes in emotion. *Psychosomatic Medicine*, 14:476–490, 1952. 23

[184] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, April 2018. 36, 44, 45, 66

[185] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 843–852, 2017. 42, 52, 168

[186] Silvan S. Tomkins and Robert McCarter. What and Where are the Primary Affects? Some Evidence for a Theory. *Perceptual and Motor Skills*, 18(1):119–158, February 1964. 18

[187] Ante Topic and Mladen Russo. Emotion recognition based on EEG feature maps through deep learning network. *Engineering Science and*

Technology, an International Journal, 24(6):1442–1454, December 2021. 75

[188] Thi-Dung Tran, Junghee Kim, Ngoc-Huynh Ho, Hyung-Jeong Yang, Sudarshan Pant, Soo-Hyung Kim, and Guee-Sang Lee. Stress Analysis with Dimensions of Valence and Arousal in the Wild. *Applied Sciences*, 11(11):5194, January 2021. 110

[189] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, March 2016. 25

[190] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. xxi, 30, 49, 112, 116, 118, 150, 176

[191] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning Factorized Multimodal Representations, May 2019. xxii, 146, 147, 148

[192] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, December 2017. 30, 45

[193] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W. Schuller. End-to-End Speech Emotion Recognition Using Deep Neural Networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093, April 2018. 112

[194] Athina Tzovara, Dominik R. Bach, Giuseppe Castegnetti, Samuel Gerster, Nicolas Hofer, Saurabh Khemka, Christoph W. Korn, Philipp C. Paulus, Boris B. Quednow, and Matthias Staib. PsPM-FR: SCR, ECG and respiration measurements in a delay fear conditioning task with visual CS and electrical US., August 2018. 66

[195] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*, January 2019. 174

[196] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30, 2017. xix, 7, 9, 30, 33, 49, 58, 61, 94, 95, 120, 122, 123

[197] Lei Wang. Feature Selection with Kernel Class Separability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1534–1546, September 2008. 44

[198] Shangfei Wang, Menghua He, Zhen Gao, Shan He, and Qiang Ji. Emotion recognition from thermal infrared images using deep Boltzmann machine. *Frontiers of Computer Science*, 8(4):609–618, August 2014. 18

[199] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, 53(3):1–34, May 2021. 51

[200] Jarosław Wasilewski and Lech Poloński. An Introduction to ECG Interpretation. In *ECG Signal Processing, Classification and Interpretation: A Comprehensive Framework of Computational Intelligence*, pages 1–20. Springer, London, 2012. 21

[201] Kuba Weimann and Tim O. F. Conrad. Transfer learning for ECG classification. *Scientific Reports*, 11(1):5251, March 2021. 175

[202] Robert S. Wilson, Christopher T. Begeny, Patricia A. Boyle, Julie A. Schneider, and David A. Bennett. Vulnerability to Stress, Anxiety, and Development of Dementia in Old Age. *The American Journal of Geriatric Psychiatry*, 19(4):327–334, April 2011. 37

[203] World Health Organization. Mental health. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response. 1

[204] World Health Organization. Mental health of older adults. https://www.who.int/news-room/fact-sheets/detail/mental-health-of-older-adults. 3

[205] Jinting Wu, Yujia Zhang, Shiying Sun, Qianzhong Li, and Xiaoguang Zhao. Generalized zero-shot emotion recognition from body gestures. *Applied Intelligence*, 52(8):8616–8634, June 2022. 18

[206] Neo Wu, Bradley Green, Xue Ben, and Shawn O'Banion. Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case. *arXiv:2001.08317 [cs, stat]*, January 2020. 49

[207] Yan Wu, Ruolei Gu, Qiwei Yang, and Yue-jia Luo. How Do Amusement, Anger and Fear Influence Heart Rate and Heart Rate Variability? *Frontiers in Neuroscience*, 13:1131, 2019. 21, 68, 175

[208] Z. Wu, X. Zhang, T. Zhi-Xuan, J. Zaki, and D. C. Ong. Attending to Emotional Narratives. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 648–654, September 2019. 30, 49, 114, 115

[209] Yanfang Xia, Filip Melinščak, and Dominik R. Bach. PsPM-VIS: SCR, ECG, respiration and eyetracker measurements in a delay fear conditioning task with visual CS and electrical US, July 2020. 66

[210] Shuo Xiao, Xiaojing Qiu, Chaogang Tang, and Zhenzhen Huang. A Spatial-Temporal ECG Emotion Recognition Model Based on Dynamic Feature Fusion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023. 25

[211] Genshen Yan, Shen Liang, Yanchun Zhang, and Fan Liu. Fusing Transformer Model with Temporal Features for ECG Heartbeat Classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 898–905, November 2019. 49

[212] H. Yang and C. Lee. An Attribute-invariant Variational Learning for Emotion Recognition Using Physiology. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1184–1188, May 2019. 40, 104

[213] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal Speech Emotion Recognition Using Audio and Text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118, December 2018. 20

[214] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-Wild: Valence and Arousal 'In-The-Wild' Challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–41, 2017. 37

[215] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A Transformer-based Framework for Multivariate Time Series Representation Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2114–2124, Virtual Event Singapore, August 2021. ACM. 57

[216] Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. A Survey on Masked Autoencoder for Self-supervised Learning in Vision and Beyond, July 2022. 174

[217] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016. 183

[218] Qiuju Zhang, Hongtao Zhang, Keming Zhou, and Le Zhang. Developing a Physiological Signal-Based, Mean Threshold and Decision-Level Fusion Algorithm (PMD) for Emotion Recognition. *Tsinghua Science and Technology*, 28(4):673–685, August 2023. 41

[219] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous Emotion Recognition using Visual-audio-linguistic Information: A Technical Report for ABAW3. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2375–2380, New Orleans, LA, USA, June 2022. IEEE. 112

[220] Tenggan Zhang, Zhaopei Huang, Ruichen Li, Jinming Zhao, and Qin Jin. Multimodal Fusion Strategies for Physiological-emotion Analysis. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 43–50, Virtual Event China, October 2021. ACM. 112, 115, 116

[221] Xiaowei Zhang, Jinyong Liu, Jian Shen, Shaojie Li, Kechen Hou, Bin Hu, Jin Gao, Tong Zhang, and Bin Hu. Emotion Recognition From Multimodal Physiological Signals Using a Regularized Deep Fusion of Kernel Machine. *IEEE Transactions on Cybernetics*, 51(9):4386–4399, September 2021. xx, 87, 88, 90

[222] Jinming Zhao, Ruichen Li, and Qin Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, Online, August 2021. Association for Computational Linguistics. 144

[223] Wei Zhou, Hao Wang, Yiling Zhang, Cheng Long, Yan Yang, and Dongjie Wang. Multi-scale Progressive Gated Transformer for Physiological Signal Classification. In *Proceedings of The 14th Asian Conference on Machine Learning*, pages 1293–1308. PMLR, April 2023. 30

[224] Jianping Zhu, Lizhen Ji, and Chengyu Liu. Heart rate variability monitoring for emotion and disorders of emotion. *Physiological Measurement*, 40(6):064004, June 2019. 22