# Multimodal Document Understanding with Unified Vision and Language Cross-Modal Learning

Souhail Bakkali

# LA ROCHELLE UNIVERSITÉ

## *École doctorale Euclide*

Laboratoire Informatique, Image, Interaction (L3i)

**THÈSE** présentée par :

**Souhail BAKKALI**

soutenance le : **14 Décembre 2022**

pour obtenir le grade de : **Docteur de La Rochelle Université**

Discipline : **Informatique**

# Multimodal Document Understanding with Unified Vision and Language Cross-Modal Learning

---

COMPOSITION DU JURY:

| | | |
|---|---|---|
| **Président du Jury** | **Mme. Elisa H. BARNEY SMITH** | Professeure, Boise State University |
| **Rapporteurs** | **M. Simone MARINAI** | Professeur, Università degli Studi di Firenze |
| | **Mme. Ergina KAVALLIERATOU** | Professeure, University Of The Aegean |
| **Examinateurs** | **M. C. V. JAWAHAR** | Professeur, IIIT Hyderabad |
| | **M. Andreas FISCHER** | Professeur, University of Applied Sciences and Arts Western Switzerland |
| **Directeur de Thèse** | **M. Mickaël COUSTATY** | Maître de conférences, La Rochelle Université |
| **Co-Encadrants** | **M. Oriol Ramos TERRADES** | Maître de Conférences, HDR, Universitat Autònoma de Barcelona |
| | **M. Marçal RUSIÑOL** (Invité) | Maître de Conférences, Universitat Autònoma de Barcelona |
| | **M. Zuheng MING** (Invité) | Maître de Conférences, Université Sorbonne Paris Nord |


Informatique
Image
Interaction
Univ. La Rochelle


La Rochelle
Université


CVC
Centre de Visió per Computador

# Abstract

The current research falls within the scope of administrative document image classification, which has been widely adopted in various document image processing applications. This thesis focuses mainly on cross-modal interactions between visual and textual information within document images, aiming for the design of an effective learning environment. The process of designing such systems involves studying the benefits of cross-modal interactions in multimodal learning. Such systems encourage cross-modal learning between visual and textual features from vision and language modalities to enhance their distribution in the common representation space. The frameworks developed were the outcome of an iterative process of analysis and synthesis between existing theories and our performed studies. In this thesis, we wish to study cross-modality learning for contextualized comprehension on document components across language and vision. The main idea is to leverage multimodal information from document images into a common semantic space. The principle consists of automatically extracting information from the content presented in the information systems (scan of documents, structured and unstructured information). Then, to understand the interactions between visual and textual data, to reorganize the research space, and to find a common semantic space to perform the required downstream applications.

This thesis focuses on advancing the research on cross-modality learning and makes contributions on four fronts: (i) to proposing a cross-modal approach with deep two-headed neural network which is capable of learning simultaneously the textual content and the visual information from scanned document images. The aim is to jointly leverage visual-language information into a common semantic representation space to automatically perform and make predictions about multimodal documents (*i.e.* the subject matter they are about); (ii) to investigating competitive strategies to address the tasks of cross-modal document classification, few-shot document classification, and content-based retrieval; (iii) to addressing data-related issues like learning when data is not annotated, by proposing a network that learns generic representations from a collection of unlabeled documents; and (iv) to exploiting few-shot learning settings when data contains only a few examples.

Keywords: Multimodal Document Understanding, Cross-Modal Document Classification, Multimodal Fusion, Few-Shot Learning, Self-Attention Mechanisms, Contrastive Learning, Deep Learning.

# Résumé

Les données papier et numériques produites par les grandes institutions publiques ou privées intègrent différents types de contenus très hétérogènes. En effet, ces contenus se présentent souvent sous diverses formes, sous forme de graphiques dans des rapports techniques, de diagrammes dans des articles scientifiques et de conceptions graphiques dans des bulletins. Effectivement, pour prendre des décisions sur des sujets d'intérêt tels que la science, les affaires, la santé, etc., l'être humain peut traiter efficacement les informations visuelles et textuelles contenues dans ces documents. Toutefois, comprendre et analyser manuellement de grandes quantités de données à partir de documents prend généralement du temps et coûte cher. En général, les données de document sont souvent présentées dans des mises en page complexes en raison des différentes manières d'organiser chaque document. Contrairement aux images générales de scènes naturelles, les documents sont très difficiles compte tenu de leurs propriétés structurelles visuelles et de leur contenu textuel hétérogène. Dans ces conditions, le développement d'outils informatiques capables de comprendre et d'extraire automatiquement des informations structurées précises à partir d'une grande variété de documents reste crucial, d'une manière qui conduit à effectuer d'importantes applications administratives et/ou commerciales. Il existe aujourd'hui plusieurs applications utilisées pour comprendre automatiquement les données des documents administratifs et commerciaux telles que : la classification des documents, la récupération de documents basée sur le contenu, la classification de documents en quelques prises de vue et le regroupement de documents. Par conséquent, la clé de la compréhension automatisée des documents réside dans l'intégration efficace des signaux provenant de multiples modalités de données. Étant donné que les documents sont nativement multimodaux, il est important de tirer parti des informations multimodales du langage et de la vision. Contrairement à d'autres formats de données tels que les images ou leur texte brut OCR, les documents combinent des informations visuelles et linguistiques, complétées par la mise en page du document. En outre, d'un point de vue pratique, de nombreuses tâches liées à la compréhension des documents sont rares. Un cadre qui peut apprendre à partir de documents non étiquetés (c.-à-d. une pré-formation), effectuer un réglage fin du modèle pour des applications de documents en aval spécifiques est plus

préféré que celui qui nécessite des données de formation entièrement annotées (c.-à-d. formés dans un mode d'apprentissage entièrement supervisé).

Le propos de cette recherche actuelle s'inscrit dans le cadre de la classification des images de documents administratifs, qui a été largement adoptée dans diverses applications de traitement d'images de documents. Cette thèse se concentre principalement sur les interactions intermodales entre les informations visuelles et textuelles dans les images de documents, visant la conception d'un environnement d'apprentissage efficace. Effectivement, le processus de conception de tels systèmes implique l'étude des avantages des interactions intermodales dans l'apprentissage multimodal. En effet, de tels systèmes encouragent l'apprentissage intermodal entre les caractéristiques visuelles et textuelles des modalités visuelles et langagières afin d'améliorer leur distribution dans l'espace de représentation commun. Encore, les cadres développés sont le résultat d'un processus itératif d'analyse et de synthèse entre les théories existantes et nos études réalisées. Notre recherche part alors du fait d'étudier l'apprentissage intermodal pour la compréhension contextualisée sur les composants du document à travers le langage et la vision. L'idée principale est de tirer parti des informations multimodales des images de documents dans un espace sémantique commun. Le principe consiste à extraire automatiquement des informations du contenu présenté dans les systèmes d'information (scan des documents, informations structurées et non structurées). Ensuite, comprendre les interactions entre données visuelles et textuelles, réorganiser l'espace de recherche, et enfin trouver un espace sémantique commun pour réaliser les applications en aval requises.

Dans l'ensemble, cette thèse se concentre sur l'avancement de la recherche sur l'apprentissage intermodalité et apporte des contributions sur quatre fronts : (i) proposer une approche intermodale avec un réseau neuronal bicéphale profond capable d'apprendre simultanément le contenu textuel et l'information visuelle de images de documents numérisés. En effet, l'objectif est d'exploiter conjointement les informations du langage visuel dans un espace de représentation sémantique commun pour effectuer et faire automatiquement des prédictions sur les documents multimodaux (c'est-à-dire le sujet dont ils traitent); (ii) étudier des stratégies concurrentielles pour s'attaquer aux tâches de classification intermodale des documents, de classification de documents en few-shot, et de récupération basée sur le

contenu; (iii) résoudre les problèmes liés aux données comme l'apprentissage lorsque les données ne sont pas annotées, en proposant un réseau qui apprend des représentations génériques à partir d'une collection de documents non étiquetés ; enfin (iv) à exploiter les paramètres d'apprentissage à quelques coups lorsque les données ne contiennent que quelques exemples.

Mots-clés : Compréhension de documents multimodaux, Classification de documents intermodaux, Fusion multimodale, Apprentissage à plusieurs reprises, Mécanismes d'auto-attention, Apprentissage contrastif, Apprentissage en profondeur.

# Acknowledgement

**Personal Acknowledgement**

First of all, I would like to express my sincere gratitude to me and myself for being such a warrior. This thesis has represented a great experience in my professional and personal life.

I want to thank all my thesis advisors, Mickaël Coustaty, Zuheng Ming, Marçal Rusiñol, and Oriol Ramos Terrades for their support and guidance in pursuing this Ph.D.

A special thanks to Zuheng. Thank you very much for the time, the support, the discussions we have always had, and most importantly the late modifications just before submitting a new paper ! I appreciate your role not only as a supervisor but, more importantly, as a mentor.

A special thanks to Mickael for giving me the chance to work on such an interesting topic, for guiding me so positively and for making me feel confident in my abilities, for responding and listening to all the requests I had, for the daily support, and also for making my Ph.D. a lot easier.

A special thanks to Marçal and Oriol. I would like to thank you for your inspiring ideas and reflections, for your very interesting comments and remarks, the meetings and conversations were vital in inspiring me to think outside the box, from multiple perspectives to form a comprehensive and objective critique. You always made me think twice ! Thanks a lot for that.

A special thanks to all the members of the jury. Thanks to the reviewers Simone Marinai and Ergina Kavallieratou for their comments and suggestions to improve my thesis manuscript. I would like to thank C. V. Jawahar, Elisa H. Barney Smith, and Andreas Fischer for agreeing to evaluate my thesis in the role of examiners.

Many thanks to the L3i laboratory and its administration for the support provided to pursue my thesis in the best possible way. More specifically to Muhammad Muzzamil Luqman who gave me the chance to come to L3i laboratory as an intern, and without

whom I would not have met Zuheng, and Mickael for my Ph.D. thesis.

Thanks to all my colleagues at the L3i laboratory who have been there to support me. My friends Mohammed Oubaidi, Yahya Samadi, Bilal Aabaichi, Mohamed Amine Es-Safhi, and Amine El Khayer for the support and encouragement throughout the process.

Last but not least, my deepest thanks to my parents Hamid and Souad, my two brothers Ilyas and Simo, and my sister Halah, for all the support you have shown me through this research, the culmination of three years of distance learning, and especially when you had to put up with my stresses and moans for the past three years of this Ph.D. Without you, I would not have been able to complete this research, and without whom I would not have made it through my engineering degree ! For my wife Saghaa, you have been amazing, for putting up with while I was sitting in the office for hours on end, and for providing guidance and a sounding board when required. I will now clear all the papers off the kitchen table as I promised !

# Contents

**Abbreviations** **xiii**

**List of Figures** **xv**

**List of Tables** **xxiii**

**1 Introduction** **1**

1.1 Motivation . . . . . 1

1.2 Problem Statement . . . . . 3

1.3 Core Multimodal Challenges . . . . . 4

1.3.1 Representation . . . . . 6

1.3.2 Alignment . . . . . 8

1.3.3 Transference . . . . . 9

1.3.4 Reasoning . . . . . 10

1.4 Background . . . . . 10

1.4.1 Language-based Representations . . . . . 11

1.4.2 Vision-based Representations . . . . . 12

1.4.3 Multimodal Representations . . . . . 18

1.5 Downstream Applications . . . . . 22

1.5.1 Datasets . . . 23
1.5.2 Document Classification . . . 24
1.5.3 Content-based Document Retrieval . . . 25
1.5.4 Few-Shot Document Classification . . . 27
1.6 Major Contributions . . . 28
1.7 List of Publications . . . 32
1.7.1 International Journals . . . 32
1.7.2 International Conferences . . . 32
1.7.3 International Workshops . . . 32

**2 Multimodal Deep Feature Fusion** **33**
2.1 Motivation . . . 33
2.2 Approach . . . 36
2.2.1 Vision Modality . . . 37
2.2.2 Language Modality . . . 38
2.3 Cross-Modal Feature Learning . . . 39
2.3.1 Visual Features . . . 39
2.3.2 Textual Features . . . 40
2.3.3 Cross-Modal Features . . . 42
2.4 Experiments and Analysis . . . 43
2.4.1 Preprocessing . . . 43
2.4.2 Implementation Details . . . 44
2.4.3 Overall Evaluation . . . 45
2.4.4 Ablation Study . . . 46
2.5 Discussion . . . 53

**3 Multimodal Deep Mutual Learning** **55**
3.1 Motivation . . . 55
3.2 Approach . . . 60
3.2.1 Vision Modality . . . 60
3.2.2 Language Modality . . . 60

3.2.3 Cross-Modal Modality . . . 61
3.2.4 Self-Attention-based Fusion Module . . . 62
3.3 Proposed Method . . . 64
3.3.1 Multimodal Mutual Learning . . . 64
3.3.2 Self-Attention-based Fusion Module . . . 67
3.4 Experimental Setup . . . 70
3.4.1 Preprocessing . . . 70
3.4.2 Implementation Details . . . 70
3.5 Experiments and Ablation Study . . . 71
3.5.1 Evaluation Protocol . . . 71
3.5.2 Intra-dataset Evaluation . . . 72
3.5.3 Intra-Dataset Confusion Matrices . . . 78
3.5.4 Inter-dataset Evaluation . . . 78
3.5.5 Inter-Dataset Confusion Matrices . . . 84
3.6 Discussion . . . 85

**4 Multimodal Document Representation Learning 89**
4.1 Motivation . . . 89
4.2 Methodology . . . 93
4.2.1 Model Architecture . . . 94
4.2.2 Cross-Modal Alignment . . . 96
4.2.3 Cross-Modal Contrastive Learning . . . 98
4.3 Experiments . . . 100
4.3.1 Pre-Training VLCDoC . . . 101
4.3.2 Fine-tuning on Multimodal Tasks . . . 101
4.3.3 Ablation Study . . . 103
4.4 Discussion . . . 109

**5 Improved Multimodal Semantic Document Representation Learning 111**
5.1 Motivation . . . 111
5.2 Method . . . 115

5.2.1 Model Architecture . . . 116
5.2.2 Pre-training Objectives . . . 117
5.3 Experiments . . . 124
5.3.1 Model Configurations . . . 124
5.3.2 Pre-Training LSRD . . . 124
5.3.3 Fine-Tuning on Multimodal Tasks . . . 125
5.3.4 Qualitative Results . . . 132
5.4 Discussion . . . 137

**6 Conclusions and Future Work** **139**
6.1 Overview . . . 139
6.2 Conclusions . . . 139
6.3 Summary of Contributions . . . 140
6.4 Future Research . . . 144

**Appendices**

**Appendix A Appendix** **149**

**Appendix B Appendix** **151**

**Bibliography** **153**

# Abbreviations

| | |
|---|---|
| AdamW | Adam Weight-Decay |
| A.E | Average Ensembling |
| AI | Artificial Intelligence |
| CNNs | Convolutional Neural Networks |
| CV | Computer Vision |
| DCNNs | Deep Convolutional Neural Networks |
| DoCVQA | Document Visual-Question Answering |
| DNNs | Deep Neural Networks |
| E.C | Equal Concatenation |
| GRU | Gated Recurrent Unit |
| GCN | Graph Convolutional Network |
| GNN | Graph Neural Network |
| LSTM | Long-Short Term Memory |
| MML | Multimodal Machine Learning |
| NLP | Natural Language Processing |
| RL | Representation Learning |
| SGD | Stochastic Gradient Descent |
| SSL | Self-Supervised Learning |
| SVM | Support Vector Machines |
| ViT | Vision Transformer |
| VQA | Visual-Question Answering |

# List of Figures

1.1 Overview of the core multimodal challenges. . . . . 5

1.2 Two types of frameworks about multimodal representation learning. (a) Joint representation aims to learn a shared semantic subspace. (b) Coordinated representation learns separated but coordinated representations for each modality under some constraints. . . . . 6

1.3 Two types of Alignment techniques to identifying cross-modal correspondences and dependencies between elements of multiple modalities, following their structure about multimodal representation learning. (a) Explicit Alignment where the goal in itself is to find the alignment (i.e. which sequence is aligned with which document image). (b) Implicit Alignment where the aim is representation taking into consideration the structure. . . 8

1.4 (a) Transfer: where both modalities will learn a representation, and from there, there will be a transfer. (b) Co-Learning: The same model gets both modalities, but at test time only one modality will be used. . . . . 9

1.5 Reasoning combines knowledge, usually through multiple intermediate steps of inference exploiting multimodal alignment and problem structure. It can be interpretable using attention weights to know where are located the most important visual features in the document, or what are the relevant meaningful words in a given text sequence. . . . . 10

1.6 Categorization of the existing document image classification methods. . . . 11

1.7 A SIFT local descriptor: red lines show the matching between keypoints of the original document (left side) and the query document (right side). . . . 13

1.8 An example of a BOVW model . . . . . 15

1.9 A Typical convolutional neural network (CNN) architecture. . . . . 16

1.10 Multimodal semantic space: Combining multiple data types such as vision and language allows us to exploit correspondences that exist between them. Different shapes are used to denote different modalities. The circle represents language feature distributions, and the triangle represents visual feature distributions. Different shapes with the same color mean that they are semantically similar in content. Different modalities reside in different feature spaces, whereby, a mapping function that transform the modalities into a common and semantic feature space is required to mitigate the heterogeneity gap, by reducing the inter-modality gap and exploring the semantic correlations. Learning this mapping still represents a complex challenge. . . . . 19

1.11 Samples of different document classes in the RVL-CDIP dataset which illustrate the low inter-class discrimination and high intra-class structural variations of document images. From left to right: *Advertisement, Budget, Email, File folder, Form, Handwritten, Invoice, Letter, Memo, News article, Presentation, Questionnaire, Resume, Scientific publication, Scientific report, Specification.* . . . . . 24

1.12 Samples of different document classes in the Tobacco-3482 dataset which illustrate the low inter-class discrimination and high intra-class structural variations of document images. From left to right: *ADVE, Email, Form, Letter, Memo, News, Note, Report, Resume, Scientific.* . . . . . 25

1.13 Overview of a multimodal deep neural network to perform cross-modal document image classification. The network is based on vision and language modalities. . . . . 25

1.14 Examples of content-based document retrieval. The first query is given from the language modality. the expected results contain relevant and semantic visual representations. Then we retrieve the category of each result as Top-k retrieved samples which belong to the same category as the language query. The second query corresponds to the query document image from the vision modality. The goal is to retrieve relevant semantic information related to the query document image. Then, the category of each result is retrieved as Top-k retrieved samples belonging to the same category as the vision query. 26

1.15 Meta-learning with an episodic task(5-way, 1-shot example). For each task, the training samples from the support set and the query samples are encoded by the embedding network. Query sample embeddings are compared with the centroid of training sample embeddings and make a further prediction. 27

2.1 Sample images from the benchmark Tobacco-3482 dataset showing the low inter-class and high intra-class of structural variations of document images. 37

2.2 The proposed cross-modal deep neural network. The $\text{NasNet}_{Large}$ model is used for the vision modality, while $\text{Bert}_{Base}$ model is used for the language modality[16, 17] . . . . . 38

2.3 Illustration of (a) LSTM and (b) gated recurrent units (GRU). (a) i, f and o are the input, forget and output gates, respectively. c and $\tilde{c}$ denote the memory cell and the new memory cell content. (b) r and z are the reset and update gates, and h and $\tilde{h}$ are the activation and the candidate activation. 41

2.4 Confusion Matrix of the Equal Concatenation fusion scheme for the proposed cross-modal feature learning network. . . . . 50

2.5 Confusion Matrix of our best cross-modal network with the superposing (*i.e.* A.E) fusion method. . . . . 51

3.1 The proposed Ensemble Self-Attention-based Mutual Learning Network (EAML [18]). . . . . 61

3.2 Sample document images and their corresponding OCR results of 9 classes of the Tobacco-3482 dataset that overlap with the RVLCDIP dataset. . . . 62

3.3 The proposed self-attention-based Fusion Module. . . . . 63

3.4 Multimodal Fusion Modality of the $\text{ML}_{Tr-KLD_{Reg}}$ method. . . . . 76

3.5 Multimodal Fusion Modality of the $\text{EAML}_{Tr-KLD_{Reg}}$ method. . . . . 77

3.6 The Precision-Recall Curves of the Inter-Dataset Evaluation of the best and the worst classes of the cross-modal modalities for the two $\text{EAML}_{Tr-KLD_{Reg}}$ and $\text{ML}_{Tr-KLD_{Reg}}$ methods. (a) illustrates the P-R curves of the best classes. (b) illustrates the P-R curves of the worst classes. . . . . 83

3.7 (a.) The Confusion Matrix of the Vision Modality of our best $\text{EAML}_{Tr-KLD_{Reg}}$) method. (b.) The Confusion Matrix of the Language Modality of our best $\text{EAML}_{Tr-KLD_{Reg}}$) method. . . . . 84

3.8 (a.) The confusion matrix of the Vision Modality of the $\text{ML}_{Tr-KLD_{Reg}}$ method. (b.) The confusion matrix of the Language Modality of the $\text{ML}_{Tr-KLD_{Reg}}$ method . . . . . 85

3.9 (a.) The confusion matrix of the Multimodal Fusion Modality of our best $\text{EAML}_{Tr-KLD_{Reg}}$ method. (b.) The confusion matrix of the Multimodal Fusion Modality of the $\text{ML}_{Tr-KLD_{Reg}}$ method. . . . . 86

4.1 Overview of the proposed cross-modal contrastive learning method. The network is composed of InterMCA and IntraMSA modules with flexible attention mechanisms to learn cross-modal representations in a cross-modal contrastive learning fashion [19] . . . . . 93

4.2 Illustration of the InterMCA and IntraMSA attention modules. The visual and textual features are transformed into query, key, and value vectors. They are jointly leveraged and are further fused to transfer attention flows between modalities to update the original features. . . . . 95

4.3 The proposed cross-modal contrastive learning objective . . . . . 99

5.1 The architecture and pre-training representation learning objectives of LSRD. LSRD is a pre-trained multimodal transformer for document understanding with unified vision and language cross-modal learning objectives. . . . . 116

5.2 Interpretation of the cross-modal projection. The visual feature $v_i$ is projected onto different text directions $l_i, l_j and l_k$. The scalar projection of $v_i$ onto the matched text sequence $l_i$ is larger than that of unmatched text sequences $l_j$ and $l_k$. . . . . 122

5.3 Vision to Vision Representative output of the retrieval process. Randomly selected Query document images are shown in the first column, and the top-5 document image retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. . . . . 133

5.4 Language to Language Representative output of the retrieval process. Randomly selected Text sequences are used as query in the first column, and the top-5 text sequence retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the queries and retrieved results for a better visualisation. . . . . 134

5.5 Vision to Language Representative output of the retrieval process. Randomly selected Query document images are shown in the first column, and the top-5 text sequence retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the retrieved text results for a better visualisation. . . . . 135

5.6 Language to Vision Representative output of the retrieval process. Randomly selected Query text sequences are shown in the first column, and the top-5 document image retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the retrieved text results for a better visualisation. . . . . 136

A.1 Vision to Vision Representative output of the retrieval process. Randomly selected Query document images are shown in the first column, and the top-5 document image retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. . . . . 149

A.2 Language to Language Representative output of the retrieval process. Randomly selected Text sequences are used as query in the first column, and the top-5 text sequence retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the queries and retrieved results for a better visualisation. . . . . 150

B.1 Vision to Language Representative output of the retrieval process. Randomly selected Query document images are shown in the first column, and the top-5 text sequence retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. The corresponding document images of the retrieved text results are shown. . . . . 151
B.2 Language to Vision Representative output of the retrieval process. Randomly selected Query text sequences are shown in the first column, and the top-5 document image retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the retrieved text results for a better visualisation. . . . . 152

# List of Tables

1.1 The Distribution of document pages over the RVL-CDIP and Tobacco-3482 datasets. . . . . 23

2.1 Overall accuracy on the Tobacco-3482 dataset versus model. E.C refers to Equal Concatenation, and S.F refers to Superposing Fusion. . . . . 45

2.2 The overall accuracy of the proposed methods with different backbones and different fusion modalities on the RVL-CDIP dataset. E.C refers to Equal Concatenation, and A.E refers to superposing Fusion. . . . . 46

2.3 Evaluation of the Vision modality against Baselines on Tobacco-3482 dataset. 47

2.4 Accuracy comparison of Language-stream state-of-the-art models on the Tobacco-3482 dataset. . . . . 47

2.5 The classification accuracy of the language streams for each class of the RVL-CDIP dataset. . . . . 48

2.6 The classification accuracy of the vision modalities for each class in RVL-CDIP dataset. . . . . 48

2.7 The classification accuracies of the cross-modal network for each class of the RVL-CDIP dataset, with the proposed fusion modalities. . . . . 49

2.8 The Recall and Precision metrics of the vision backbones of the most relevant classes in the RVL-CDIP dataset. . . . . 52
2.9 The Recall and Precision metrics of the vision backbones of the most relevant classes of the RVL-CDIP dataset. . . . . 53

3.1 The overall classification accuracy of our best EAML$_{Tr-KLD_{Reg}}$ method against baseline methods on the RVL-CDIP dataset. . . . . 73
3.2 The overall classification accuracy(Acc.), recall(R.), precision(Pr.) metrics of the proposed approaches on the RVL-CDIP dataset. IL, ML$_{KLD}$, ML$_{Tr-KLD_{Reg}}$, and EAML$_{Tr-KLD_{Reg}}$ denote Independent Learning, Mutual Learning with the standard KLD, Mutual Learning with the truncated-KLD, and Ensemble self-attention-based Mutual Learning with the truncated-KLD respectively. . . . . 74
3.3 The overall classification accuracy(Acc.), recall(R.), precision(Pr.) metrics of the proposed approaches on the Tobacco-3482 dataset. IL, ML$_{KLD}$, ML$_{Tr-KLD_{Reg}}$, and EAML$_{Tr-KLD_{Reg}}$ denote Independent Learning, Mutual Learning with the standard KLD, Mutual Learning with the truncated-KLD, and Ensemble self-attention-based Mutual Learning with the truncated-KLD respectively. . . . . 78
3.4 The overall classification accuracy of the proposed approaches against baseline methods on the Tobacco-3482 dataset. . . . . 79
3.5 The Inter-Dataset Evaluation results of the Mutual Learning ML$_{Tr-KLD_{Reg}}$ method on the Tobacco-3482 dataset. . . . . 80
3.6 The Inter-Dataset Evaluation results of the Ensemble Self-Attention Mutual Learning (EAML$_{Tr-KLD_{Reg}}$) approach on the Tobacco-3482 dataset. . . . . 80
3.7 The Inter-Dataset Evaluation results of the Ensemble Self-Attention Mutual Learning (EAML$_{Tr-KLD_{Reg}}$) approach on the RVL-CDIP dataset. . . . . 81
3.8 The average precision (AP) scores of the inter-dataset evaluation of the ML$_{Tr-KLD_{Reg}}$ and the EAML$_{Tr-KLD_{Reg}}$ for the multimodal Fusion modality on the Tobacco-3482 dataset. . . . . 82

4.1 Ablation study on VLCDoC on cross-modality attention components, pre-trained on Tobacco dataset. . . . . 104
4.2 Top-1 accuracy (%) comparison results of our proposed cross-modal contrastive learning loss against the standard supervised contrastive learning (SCL) loss on the Tobacco dataset. . . . . 105
4.3 Cross-dataset test on datasets with different size and document types. Tobacco-3482, RVL-CDIP, Tobacco-3482 → RVL-CDIP denotes pre-train on the Tobacco-3482, fine-tune and test on RVL-CDIP. . . . . 106
4.4 Top-1 accuracy (%) comparison results of different document classification methods evaluated on the of RVL-CDIP dataset. V+L denotes vision+language modalities. . . . . 107
4.5 Intra-Dataset and Inter-dataset evaluation on the Few-shot document classification setting. The best embedding network is pre-trained on RVL-CDIP dataset, then tested on Tobacco-3482 dataset. All accuracy results are averaged over 600 test episodes and are reported with 95% confidence intervals. 109

5.1 Top-1 accuracy (%) comparison results of different document classification methods evaluated on the of RVL-CDIP dataset. V, T, and L denote Vision, Text, and Layout modalities. . . . . 125
5.2 Few-shot classification accuracy results on the test set of the RVL-CDIP dataset. All accuracy results are averaged over 600 test episodes and are reported with 95% confidence intervals. MLP denotes the projector used on top of the vision and language modalities to perform (VLN-NCLR) pre-training objective. ME denotes the multimodal encoder used on top of the vision and language modalities to perform the vision-language matching pre-training objective (VLM). **('+', '-')** indicate results w/wo meta-learning. 127
5.3 Quantitative evaluation results of Intra-Modal and Inter-Modal Content-based retrieval on RVLCDIP 40K test set in terms of Recall@K(R@K). . . 130
5.4 Effects of sequence length on content-based document retrieval. . . . . 132

# CHAPTER 1

# Introduction

*Multimodal presentations have an inherent critical potential to the extent that we learn how to use the images to deconstruct the viewpoint of the text, and the text to subvert the naturalness of the image.*

*– Jay Lemke*

## 1.1 Motivation

Multimodal Machine Learning (MML) has seen increased attention lately and has been considered as an active multi-disciplinary research field. MML addresses some of the original goals of artificial intelligence (AI) which have been already incorporated in many domains by integrating and modeling multiple sensory input modalities including linguistic, visual, and layout information. This research field brings various challenges for multimodal researchers given the heterogeneity of document data and the contingency often found between its different modalities. Intuitively, the multimodality of documents require multimodal reasoning over multimodal inputs, where data related to the same topic of interest tend to appear together. These multimodal inputs (e.g., visual, textual,

and layout) within document images are presented in a diverse set of sources such as handwritten text, tables, forms, figures, multi-column layouts, plain text, curved text, and exotic fonts, etc. As humans, we regularly extract information from illustrations in document data like advertisement, scientific publications, and articles; parse graphs and charts to make decisions; and allow informational data to influence our opinions regarding the type and/or the category of these documents, as the visuals burn into our memory. Meanwhile, understanding documents visually encounters the problem of low inter-class discrimination, and high intra-class structural variations between the different categories of document data. In a general way, visual data can be more telling (a picture is worth a thousand words). However, some documents contain abundant visual information such as reports, and scholarly articles, in which case a stronger emphasis on the semantic meaning of language is more helpful. Therefore, handling the semantic and stylistic variability in documents is challenging to computational models that are trained mostly on natural images. Furthermore, multimodal reasoning allows one to integrate information from language and vision modalities, to reason about the structure of the documents (e.g., how the accompanying figures support the text), and to gather the relevant semantic information from the text corpus (e.g., how to distinguish between a letter and an email), to finally gather the most important information within the common representation space for decision-making. Hence, multimodal reasoning has been the defacto for many document understanding research projects which fall at the interface of Computer Vision (CV), and Natural Language Processing (NLP) (if an image is worth a thousand words, then a multimodal document is worth a thousand concepts).

This thesis is mainly centered on cross-modality learning, focusing more on the most frequent and principal modalities studied in the state-of-the-art, which are the vision and the language ones. The first parts of this thesis address the document classification problem in a fully supervised learning fashion. At first, frameworks that project uni-modal representations together into a joint multimodal representation space are proposed. Joint representations are mostly used in tasks where multimodal data is present during both the training and inference steps. The simplest example of a joint representation is an early fusion methodology such as dot product, concatenation, and average ensembling of

individual modality features. Second, an alternative to a joint multimodal representation is a coordinated representation which has been explored in the following parts of this thesis. Instead of projecting the vision and language modalities into a joint representation space, we learn vision-language representations by coordinating them through a mutual learning constraint. We start our discussion with coordinated representations that mimic the probability distributions of each modality, moving on to coordinated representations that enforce similarity between representations, to finally address more powerful coordinated representation constraints. The latter focuses more on enhancing the structure of the resulting representative space in a self-supervised learning fashion, where we introduce two novel downstream applications (*i.e.* vision-language few-shot document classification, and vision-language content-based document retrieval) that were not established before in the document understanding literature. We evaluate our proposed strategies on publicly available benchmark document datasets, compared to the most recent state-of-the-art studies related to document understanding.

The following section provides an overview of the main challenges of multimodal document understanding (*e.g.* a document can be either a scanned image or plain text). Later on, we refer to multimodal document understanding as the ability of a system to use multiple sensory modalities (*i.e.* multiple data inputs: vision and language) to perform a desired task. In contrast, we refer to cross-modal learning as the ability of that system to use and learn information from different modalities to improve the performance of the system (*i.e.* a scientific publication can be categorized by its visual spatial properties and by semantic language information).

## 1.2 Problem Statement

In general terms, multimodal learning is more related to sensory modalities like the sound, the speech, the touch, etc. A modality refers to a certain type of information and/or the representation format in which information is stored. The word modality is mostly associated with sensory modalities which are one of the primary forms of sensation, like vision or touch, considered as channels of communication. Also, thinking of multimodality

engender thinking of multi-disciplinary. This comes from many different fields all together to be combined/involved in an approach to a topic or a problem, in a sense that almost Artificial Intelligence (AI) is coming together: there are the vision, the language, and the aspect of learning cross-modal knowledge. Specifically, the core of this thesis is about multimodal document understanding, based on the vision and the language, as two of the building blocks of our application on document data. In order for AI to make progress in understanding the world around us, it needs to be able to interpret and reason about multimodal messages. Multimodal document understanding aims to build models that can process and relate information from multiple modalities related to a phenomenon. This can provide different perspectives which enable a system to:

- Learn complementary and additional information to transfer knowledge from each modality to another in a collaborative learning fashion, in contrast to dealing with just uni-modal modalities.
- Discover patterns or changes that are only visible when two or multiple modalities are studied.
- Capture correspondences between modalities and gaining an in-depth understanding of a natural phenomenon.

Therefore, it is crucial to develop systems that may lead to some enlightenment about the world around us, by thinking of systems that learn from multimodal sources.

## 1.3 Core Multimodal Challenges

The research field of multimodal document understanding brings some unique challenges given the heterogeneity of document data. The core challenge for many problems but also for multimodal document understanding is how to bring vision and language together. The first level is very important as everyone nowadays uses deep neural networks (DNNs) to understand the challenges of representation learning. There are some key core challenges that are related to multimodal representation learning: Alignment, which is a very multimodal key with the goal to identify relations between elements from two or multiple

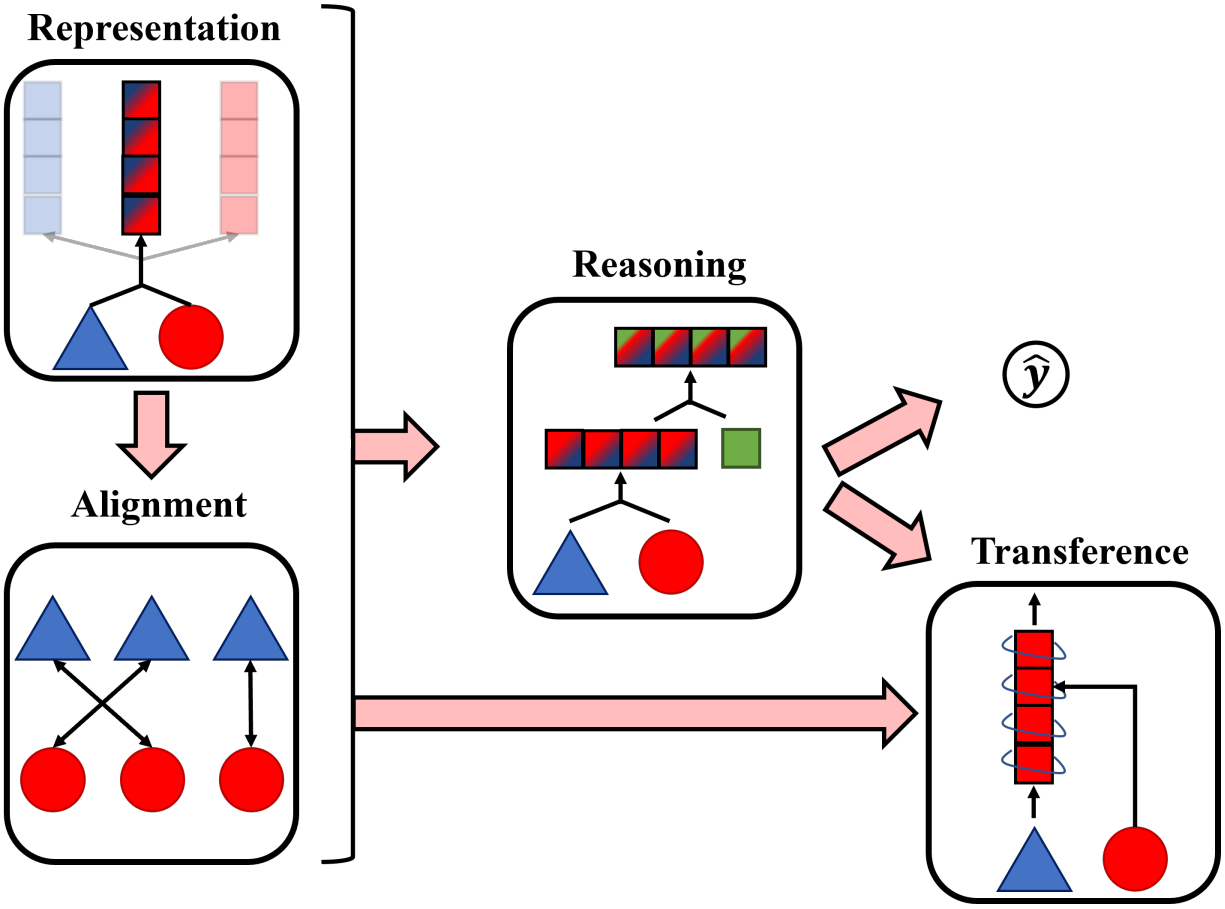



Figure 1.1: Overview of the core multimodal challenges.

different modalities. As humans, we could learn a representation. For example, when we say "I like it!" with a happy face, or when we are tense or surprised, the representation will encode that similarity at some level. Today, we are a lot closer to that, because as humans, we are able to learn joint representations where we see an object as a visual representation, and we see some language associated with this object as a language representation. We observe that we have some kind of paired data, and so a joint representation that allows to learn this one space where both of them will coexist together. In the 2010-2011 era, this sounded impossible. However, we have seen a lot more of that in the recent past. We got to see these kind of joint representations as a very important milestone. In fact, Representation Learning (RL) can be defined as learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy. For example, when we have multiple documents from the same category (*e.g.* Scientific publications,

emails, etc.), these types of documents share the same visual spatial information that we want to take advantage of in order to be more efficient and more robust. Meanwhile, we also want to do complementarity, like when two things are not sufficient by themselves and we want to bring them together. One of the greatest challenges of multimodal data is to summarize the information from multiple modalities (or views) in a way that complementary information is used as a conglomerate while filtering out the redundant parts of the modalities. Due to the heterogeneity of the data, some challenges naturally spring up including different kinds of noise, alignment of modalities (or views), etc. To sum up, we explore four challenges which are: representation, alignment, transference, and finally reasoning [20] as depicted in the Figure 1.1.

### 1.3.1 Representation

Good representations are important for the performance of MML models. This first core challenge is concerned with how to represent and summarize multimodal data, by either fusing or coordinating them.

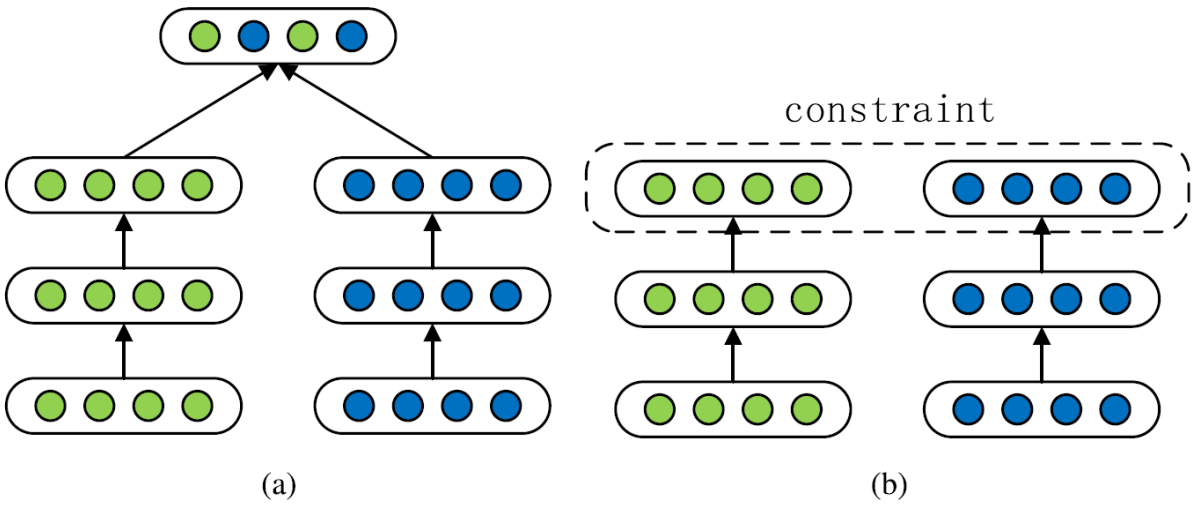



Figure 1.2: Two types of frameworks about multimodal representation learning. (a) Joint representation aims to learn a shared semantic subspace. (b) Coordinated representation learns separated but coordinated representations for each modality under some constraints.

### Fusion

Fusion involves projecting all the different modalities to a common representation space while preserving information from the given modalities. In this type of representation learning, input data from all modalities is required at the training and inference steps which can potentially be hard while missing some kind of input data. In our study, we propose a case-study model which can fuse different views of a modality at each time-step and finally use the joint representation to complete the required downstream tasks as in Figure 1.2(a). This task can be performed in a late, early, intermediate, or attention-based fusion approach [53, 91].

### Coordination

Instead of bringing everything together, we bring each one of the language and vision modalities, having their one representation space. The coordination involves projecting all the modalities to their space coordinated using a constraint. The coordination should be seen as a spectrum; At one end, the coordination can be so strong that the representations are equal, forcing the language representation to be equal to the visual representation. At this point, it is mostly a joint representation. At the other end, the representations are separate, so we don't coordinate at all. One example is to say, instead of making them equal, we make them correlated (*i.e.* it is not as much as equal but close). Another example is to say we are going to bring together only a subset of each representation; there are some items that we want to be very close to each other, and for the rest we will let each modality separate and let them be themselves (see Figure 1.2(b)). This kind of approaches is more useful for modalities which are fundamentally very different and might not work well in a joint space. Due to the variety of modalities in nature, Coordinated Representations have a huge advantage over Joint Representations which gives us reason to believe that the coordination using constraints is the way to go in the field of multimodal representation learning [66, 206].

### 1.3.2 Alignment

One thing that is core to multimodal learning is alignment, like synchrony, where we want to be able to align speech and reading as an example. Alignment is defined as identifying cross-modal interactions and the direct relations between (sub)elements from two or multiple different modalities, building from the data structure [37]. Alignment can be differentiated as Explicit and Implicit:

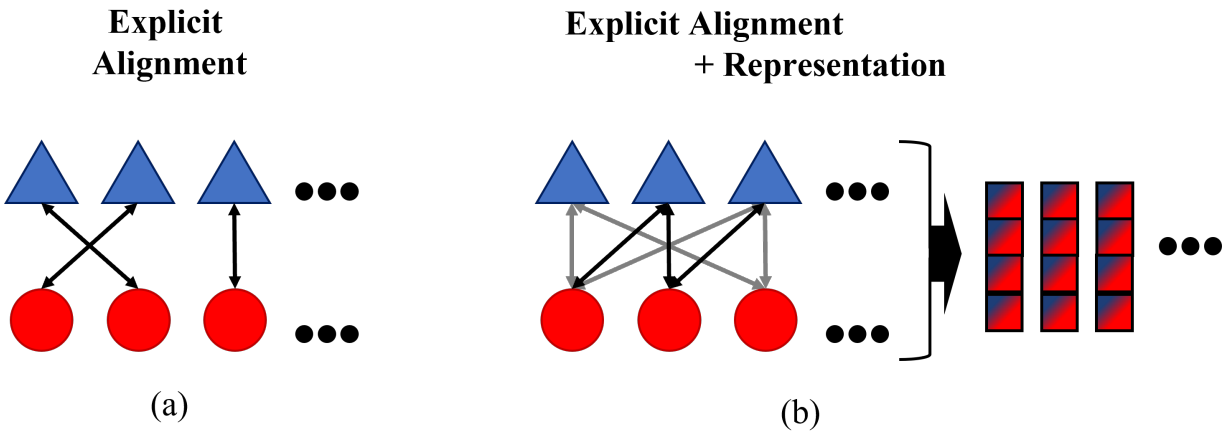



Figure 1.3: Two types of Alignment techniques to identifying cross-modal correspondences and dependencies between elements of multiple modalities, following their structure about multimodal representation learning. (a) Explicit Alignment where the goal in itself is to find the alignment (i.e. which sequence is aligned with which document image). (b) Implicit Alignment where the aim is representation taking into consideration the structure.

#### Explicit Alignment

Explicit Alignment is defined as taking advantage of how each modality has its internal structure. Some modalities might be temporal, spatial, or hierarchical, etc. Within a specific modality, a document image has in it multiple elements, and each one of them are linked somewhat. As such, explicit alignment enables not only linking elements within a modality, but more interestingly between modalities, being able to see which element from one modality connects with the other element from the other modality (see Figure 1.3(a)). The sub-challenge here is to directly find correspondences between elements from different modalities (ex. which sequences align the most with which document image).

### Implicit Alignment

In the world of deep learning, the alignment task is often defined as a sub-task, a latent process where the real task is representation where we take into consideration the structure. One popular architecture that is mostly used nowadays is the transformer-based architecture; this type of architecture applies implicit alignment and they often end up being fully connected, aiming to look at what are the relevant elements between modalities and then learn new representations from that (see Figure 1.3(b)).

## 1.3.3 Transference

Transference is defined as transferring knowledge between two or different modalities, usually to help the target modality which may be noisy or have limited resources [138, 203]. The idea behind is having one modality which doesn't have as much data or noisy, and the other modality will come to help. There are two sub-challenges of transference: Transfer and Co-learning.

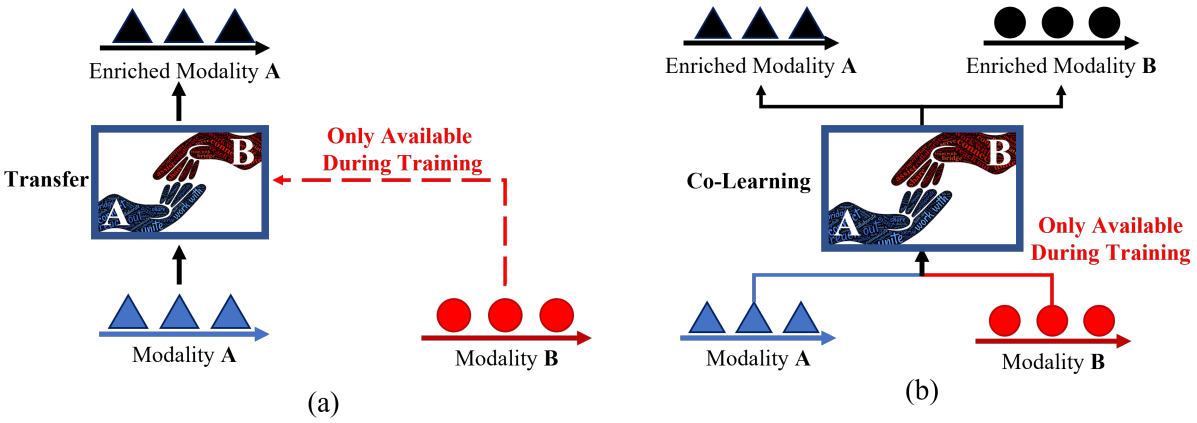



Figure 1.4: (a) Transfer: where both modalities will learn a representation, and from there, there will be a transfer. (b) Co-Learning: The same model gets both modalities, but at test time only one modality will be used.

### Transfer

Transfer is where both modalities will learn a representation and then from there there will be a transfer (see Figure 1.4(a)).

### Co-Learning

In co-learning, the same model gets both modalities as input, but at test time, only one modality will be used (see Figure 1.4(b)).

### 1.3.4 Reasoning

Another core challenge in multimodal learning is to try to not just look at lower levels, but also to think about how do we combine knowledge, usually through multiple steps of inference to exploit the alignment and the problem structure [116]. Reasoning goes beyond a local representation or a representation with alignment (see Figure 1.5).

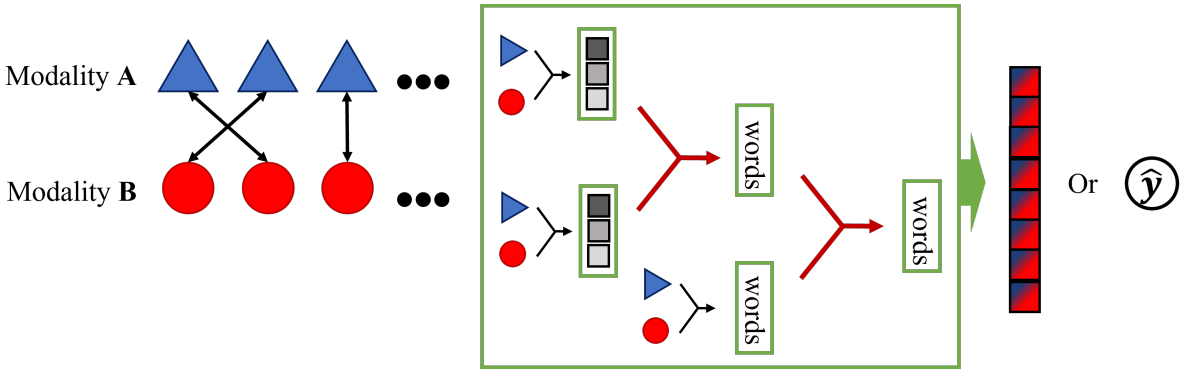



Figure 1.5: Reasoning combines knowledge, usually through multiple intermediate steps of inference exploiting multimodal alignment and problem structure. It can be interpretable using attention weights to know where are located the most important visual features in the document, or what are the relevant meaningful words in a given text sequence.

## 1.4 Background

One important part of any problem in multimodal document understanding is related to the representation of the data involved. Representation learning aims to find representations of raw and unstructured data as useful information to perform tasks such as classification or prediction. Good representations are important for the performance of machine learning models. While the development of uni-modal representations has been extensively studied, multimodal representations still represent a challenge. In this section,

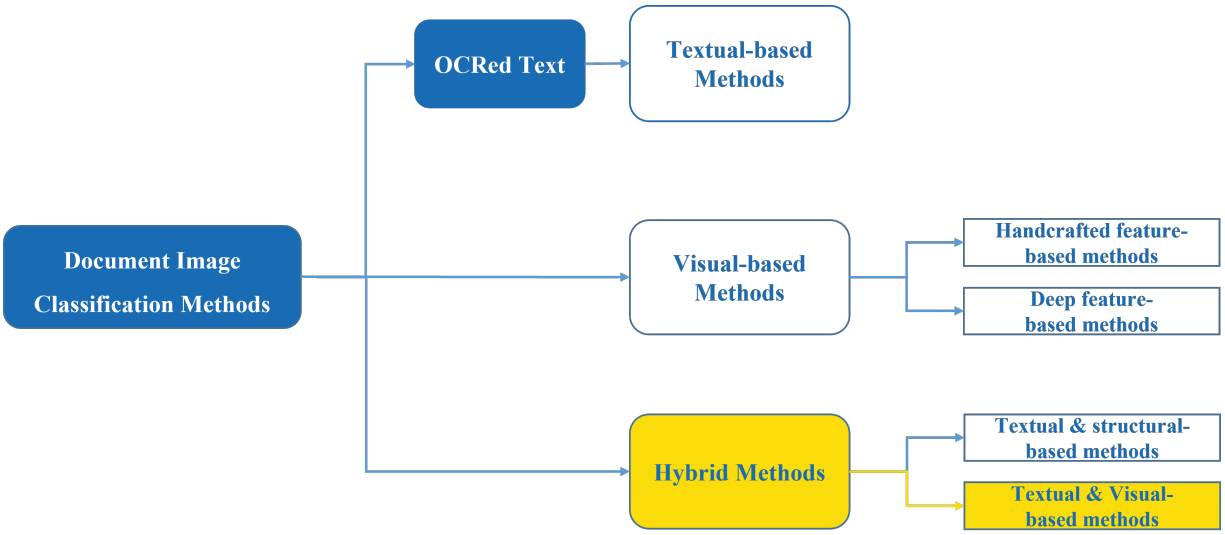



Figure 1.6: Categorization of the existing document image classification methods.

we present the evolution and description from uni-modal to multimodal representations as illustrated in Figure 1.6. We perform a general review of the main representation models for language and vision separately, ending with the main approaches for multimodal learning for these two modalities. We also present the state-of-the-art of tasks addressed in this thesis, with a summary of our proposed strategies that will be presented in the following chapters.

### 1.4.1 Language-based Representations

Regarding the language-based representations, they are extracted from the textual content generated from an Optical Character Recognition (OCR) [73] engine. Then, the textual content is used to perform the desired downstream task. In many natural language processing (NLP) tasks, the representation of words has drawn significant attention. The development of static word embeddings such as Word2Vec, Glove [118, 135], to contextualized dynamic word embeddings such as ELMO, Fasttext, XLNet, and Bert [41, 119, 137, 190] have made a huge progress to address the polysemy problem and the semantic aspect of words. Meanwhile, several approaches handled the task of document image classification by performing OCR techniques [50, 92, 120]. The task of document image classification is then transformed into text classification [202]. Yang *et al.* [189] combined

generated text features with visual features in a fully convolutional neural network. Also, [13, 38] experimented with shallow Bag-of-Words (BoW) [40] along visual features in a two-modality classifier. Moreover, similar to our approach, Lai *et al.* [92] presented a hybrid approach to extract contextual information using a RNN-CNN.

### 1.4.2 Vision-based Representations

Over the past few years, a variety of research studies have been proposed for document understanding. Due to the different manners of organizing each document, document images might be classified based on their heterogeneous visual structural properties and/or their textual content. The visual appearances of document images can be divided into two subcategories which are: handcrafted feature-based methods, and deep feature-based methods. We will elaborate these two types of methods in the following sub-sections.

#### Handcrafted Feature-based Methods

Handcrafted feature extraction methods are used to extract features from the document images, which are then utilized to train machine learning classifiers. Different handcrafted feature-based methods have been employed to perform the task of document image classification. Some of the handcrafted features used for document data include (1) local and global descriptors, (2) bags-of-visual-words, and (3) other miscellaneous methods.

**Local and Global Descriptors.** Local features describe the document image patches (key points in the document image) of an object, and represent the patterns in a specific region which differs from its immediate neighborhood. There exist a variety of local descriptors such as Scale Invariant Feature Transform (SIFT) [113], Principal Component Analysis with Scale Invariant Feature Transform (PCA-SIFT) [78], Speeded Up Robust Features (SURF) [22], Gradient Location and Orientation Histogram (GLOH) [117], Shape context [23], and so on. These descriptors have been employed widely in computer vision tasks such as image classification [40, 154], object tracking [208], etc. Earlier attempts on document understanding have applied local descriptors to classify document images. In [96], Phuong *et al.* proposed a logo spotting model based on the matching key-points extracted from the document images and a given set of logos using SIFT. Specifically, local

features are used to describe the logo and document images. Then, the detected key-points between the document image and each logo are matched based on their SIFT descriptors using its two nearest neighbors in the SIFT feature space [113]. Figure 1.7 illustrates an example of SIFT local descriptor. Also, in [97], improve the problem of the unmatched key-point pairs in [96] by filtering the incorrectly matched key-points based on filter by homography. Although local descriptors are robust to image distortions, they encounter the problem of having many local descriptors when computed on document images, which leads to inconsistent classifiers. As for the global descriptors, they describe the image as a whole to generalize the entire object, such as intensity, textures, and color histograms. Global descriptors include histogram-oriented gradients (HOG) [6] are generally used in image retrieval [68], object detection [122] and image classification [84].

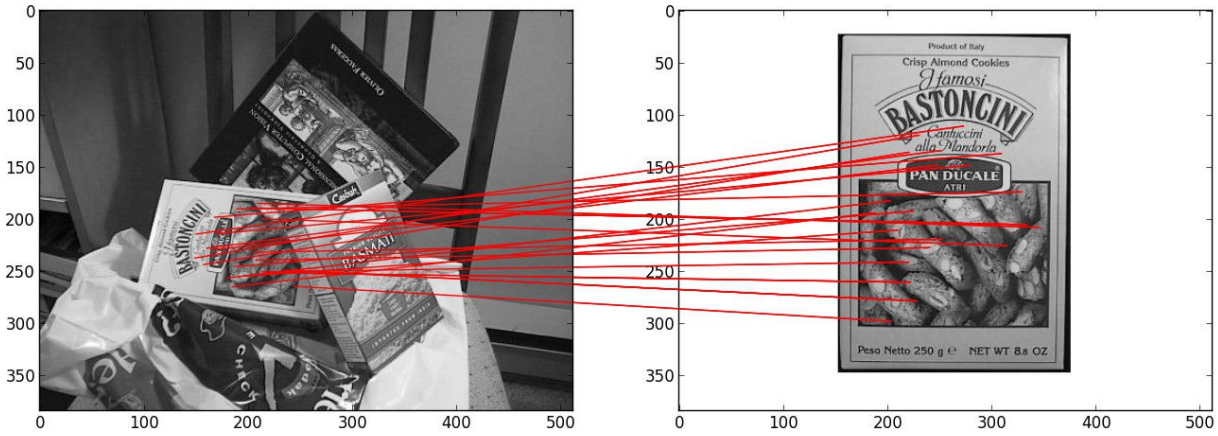



Figure 1.7: A SIFT local descriptor: red lines show the matching between keypoints of the original document (left side) and the query document (right side).

**Bag-Of-Visual-Words (BOVW) and Fisher Vectors.** The Bag-of-Visual-Words (BOVW) [159] is commonly used in image classification [40, 77]. The general idea is adapted from information retrieval and NLP's Bag-Of-words (BOW) [205]. In Bag-Of-Words (BOW), the number of each word appearing in a document is counted, where the frequency of each word is used to know the keywords of the document, and then, a frequency histogram is made from it. In the text domain context, a document is treated as a Bag-Of-Words (BOW). However, in the vision domain, a document image is represented as a set of features, which consist of key-points and descriptors (the description of the

key-points). Vocabularies are constructed from the extracted key-points and descriptors to represent each document image as a frequency histogram of features in the document image. From the frequency histogram, the category of the document image can be then predicted. Figure 1.8 illustrates an example of a BOVW model, which is composed of two steps: vocabulary learning and representation generating. In vocabulary learning, given a training set of document images, the descriptors are first extracted with a local descriptor (*e.g.* SIFT). Then, a clustering approach is employed such as K-means [74] to group the descriptors extracted into $K$ clusters, where each cluster is considered as a visual word. Afterwards, a visual vocabulary consisting of $K$ words is built to finally represent document images as vectors. The BOVW model has been applied to document image classification. In [149], Rusinol *et al.* employed BOVW for document classification using logo spotting. Also, Kumar *et al.* [88–90] learned the characteristics of document images by extracting patches of the document images both horizontally and vertically. Further, the BOVW model is employed to generate visual representations for each extracted region as in [95]. Then, a random forest classifier is performed for the task of document image classification. The Fisher vector representations [136] are considered as an extension of the BOVW model. They extract a set of local patch descriptors to encode them in a high dimensional feature vector. It has been applied to classify document images [34, 35].

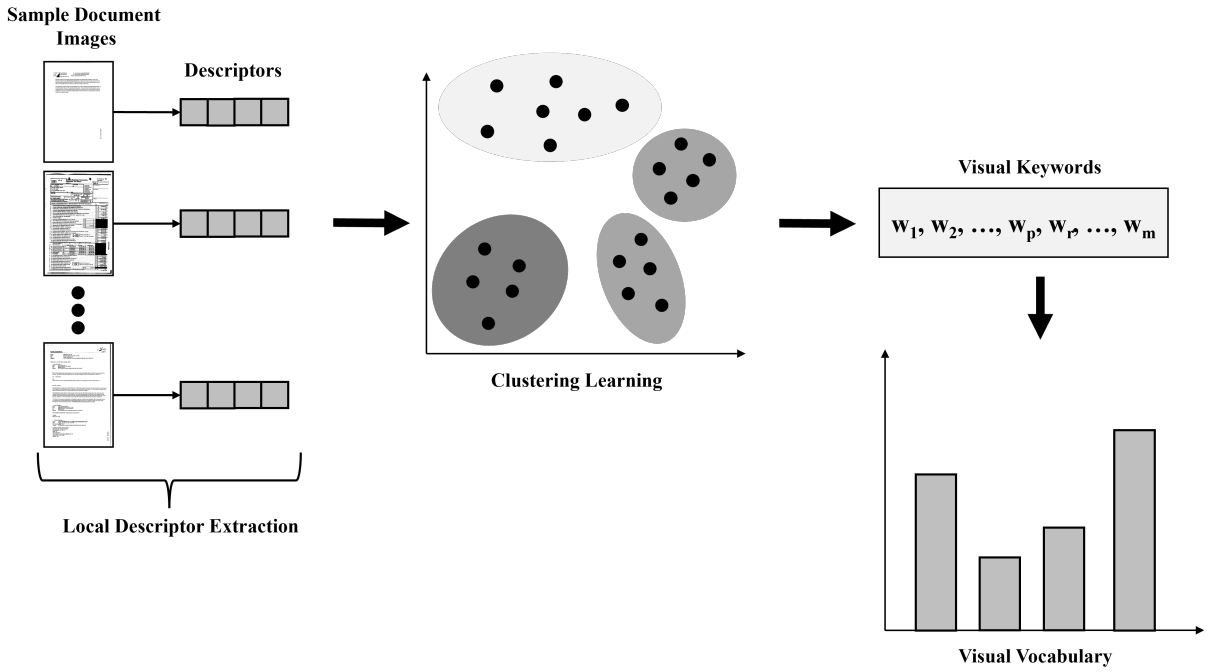



Figure 1.8: An example of a BOVW model

**Miscellaneous Methods.** There exist other handcrafted feature-based methods that have been applied in the document image classification task. For instance, Sarkar *et al.* [151] used Viola-Jones based features to represent document forms, followed by a latent conditional independence model to perform classification. Besides, Usilin *et al.* [171] proposed a new Viola-Jones based method to classify documents and to detect the placement and orientation of documents within an image. Some other works employed histograms and binarization methods to classify documents. Gordo *et al.* [57] represented document images using binarized runlength histograms followed by a 1-NN classifier to perform classification. Also, Reddy *et al.* [83] applied binarization on document images along pixel density, followed by K-means clustering and adaptive boosting methods for form classification.

### Deep Feature-based Methods

Image data is represented as a two-dimensional grid of pixels, be it monochromatic or in color. Accordingly each pixel corresponds to one or multiple numerical values respectively. The advancements in computer vision with deep learning have been constructed and per-

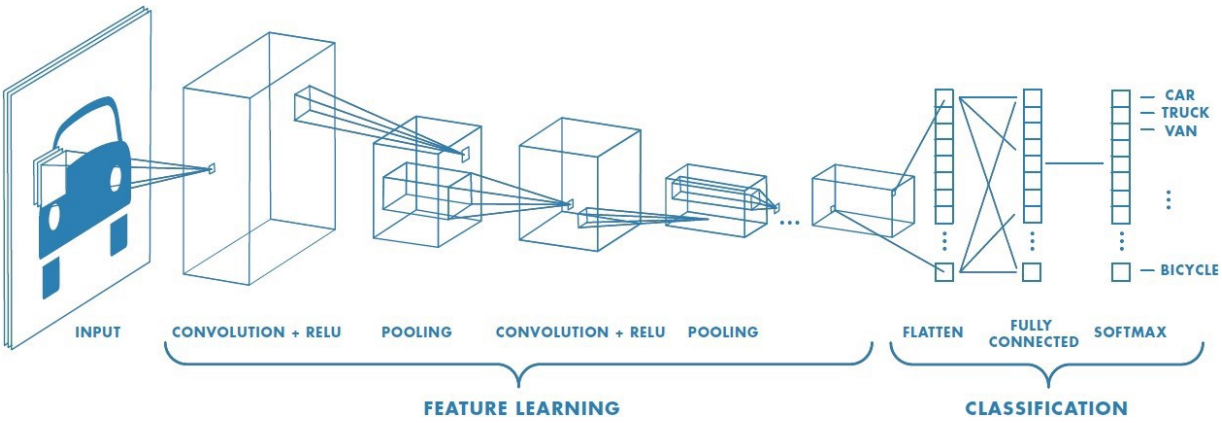



Figure 1.9: A Typical convolutional neural network (CNN) architecture.

fected with time, primarily over one particular algorithm — a Convolutional Neural Networks (CNN) [99]. A CNN is a deep learning algorithm which can take in an input image, assign learnable weights and biases to various aspects and/or objects in the image, and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms. While in primitive methods described above filters are hand-crafted, CNNs have the ability to learn these filters/characteristics of images with enough training time. With the great success of AlexNet architecture in the ImageNet Large-Scale Visual Recognition Challenge 2012 (ILSVRC2012) contest [87], CNNs have been extensively applied in various computer vision tasks [80, 167, 199]. Figure 1.9 illustrates a typical CNN architecture. The first component represents feature learning which are able to successfully capture the spatial and temporal dependencies in an image through the application of relevant filters. The features extracted are called deep features. The other component is fully connected. It takes care of the classification using the obtained deep features. The architecture performs a better fitting to the image dataset due to the reduction in the number of parameters involved and re-usability of weights. In other words, the network can be trained to understand the sophistication of the image even better. For instance, CNNs have been extensively utilized in document understanding approaches, and more specifically, for document image classification, which may be either trained document image classifiers in an end-to-end manner [1, 3, 4, 61, 76, 165, 166, 187], or used as an off-the-shelf feature extractor [35, 36, 85, 148, 152]. Training convolu-

tional neural networks (CNN) in an end-to-end fashion was firstly adopted by Kang *et al.* [76]. They proposed a shallow CNN representation to classify documents. they showed that CNNs outperform handcrafted feature-based methods, which indicates the potential of deep features. The AlexNet architecture is employed in [3] for document classification. Despite the large differences between document images and images in the ImageNet dataset [39], they showed that pretrained weights perform better than a random weight initialization. Moreover, Tensemeyer *et al.* [166] conducted an exhaustive investigation of numerous factors that have an impact of the classification performance of convolutional neural networks (CNNs). These factors include document image size, aspect ratio preservation, training set size, data augmentation, etc. Furthermore, Afzal *et al.* [4] trained four convolutional networks including AlexNet, VGG-16 [158], GoogLeNet [165], and ResNet-50 [63] to perform document classification. They also investigate the performance of CNNs with and without the pretrained ImageNet weights, and confirm their effectiveness in improving the classification performance compared to random weigh initialization. Likewise, several different deep CNNs such as Inception-ResNet-v2 [164], DenseNet [69], and ResNeXt [183] have been proposed and proved to be effective for document image classification on the large-scale RVL-CDIP[1] and the low-scale Tobacco-3482 datasets. In [85], Kolsch *et al.* proposed to classify document images by replacing the last fully connected layer in the AlexNet architecture with the Extreme Learning Machines as in [70, 71]

Therefore, employing convolutional neural networks (CNNs) as an off-the-shelf feature extractor has also been studied, where feature extraction and classifier learning are conducted in an integrated fashion. The obtained pretrained features are then passed to a classifier to perform the final downstream task (*i.e.* document image classification). In [148], Roy *et al.* divided document images into five separate sections: header, footer, left body, right body, and the whole document image. After training each CNN on each different section, generalized stacking is employed to combine the five normalized outputs by training a meta-classifier (*i.e.* SVM in this work). In [36], Das *et al.* utilized two levels of transfer learning. On the one hand, inter-domain transfer learning is used for training the VGG-16 network for the whole document image with ImageNet pretrained weights.

[1] https://www.cs.cmu.edu/~aharley/rvl-cdip/

On the other hand, intra-domain transfer learning is used for training the VGG-16 models for four sections (*i.e.* header, footer, left body, and right body) with the latter VGG-16 pretrained model on the whole document image. Finally, a stacked generalization scheme is used to combined the predictions of the different pretrained VGG-16 models as in [148].

Generally speaking, training convolutional neural networks in an end-to-end fashion works well when a large amount of training samples is available. Nevertheless, employing convolutional neural networks as off-the-shelf feature extractors is more helpful when training data is limited. Thus, when pretrained, a suitable classifier is crucial to achieve compelling performance.

### 1.4.3 Multimodal Representations

In this subsection, we present some multimodal methods that have been proposed recently, which combine textual features with either visual features, or structural features. The common pipeline of multimodal methods involves two streams: textual stream, visual stream. For the textual stream, the text is first extracted from the document image based on an OCR engine. then, a text classifier is trained for the textual stream, and an image classifier is trained for the image stream. Finally, the two streams are fused to determine the class of the document image (see Figure 1.10). In an attempt to fuse the two streams, different fusion strategies can be employed, *e.g.* early fusion, late fusion, and middle fusion.

#### Joint Representations

As stated before, documents are natively multimodal. Multimodal learning for computer vision and natural language processing has been widely used for image and text level understanding problems such as text document image-based classification, visual question answering [191, 207], image captioning [8] and image-text matching [102]. Most multimodal fusion and attention learning methods require multimodal reasoning over multimodal inputs that are represented into a common space, where data related to the same topic of interest tend to appear together. For the multimodal fusion methods, earlier attempts used naive concatenation, element-wise multiplication, and/or ensemble methods for mul-

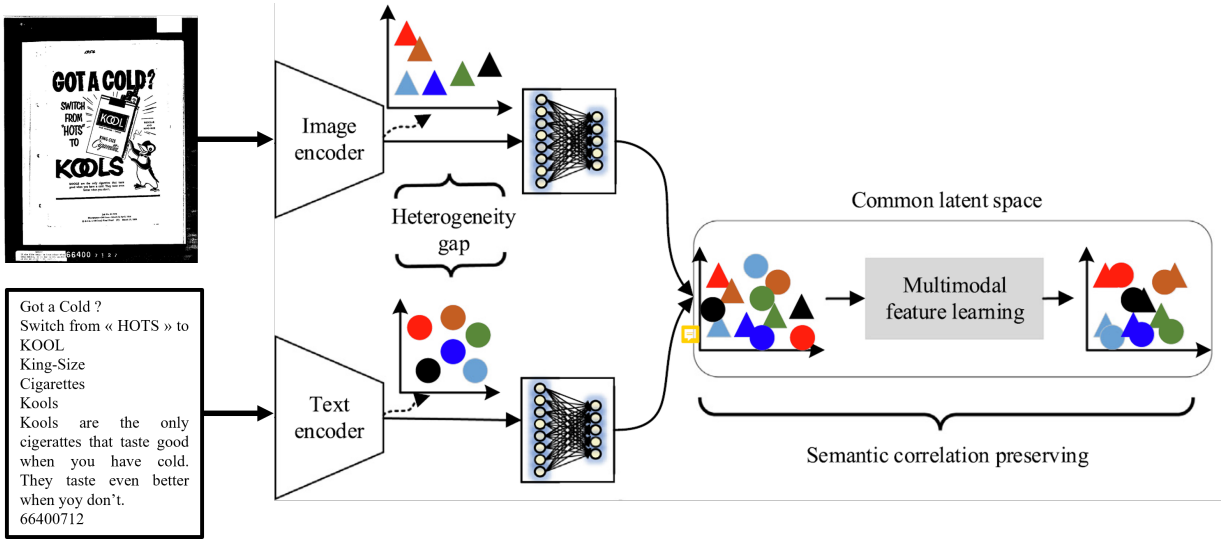



Figure 1.10: Multimodal semantic space: Combining multiple data types such as vision and language allows us to exploit correspondences that exist between them. Different shapes are used to denote different modalities. The circle represents language feature distributions, and the triangle represents visual feature distributions. Different shapes with the same color mean that they are semantically similar in content. Different modalities reside in different feature spaces, whereby, a mapping function that transform the modalities into a common and semantic feature space is required to mitigate the heterogeneity gap, by reducing the inter-modality gap and exploring the semantic correlations. Learning this mapping still represents a complex challenge.

timodal features [11, 12, 52, 109, 157, 188, 189, 196, 200]. Noce *et al.* [126] proposed an approach that combines OCR and NLP algorithms to extract and manipulate relevant text concepts from document images, which are visually embedded within each document image to improve the classification results of a convolutional neural network. Fukui *et al.* [51] proposed a multimodal compact bi-linear pooling to efficiently and expressively combine multimodal features. Yang *et al.* [189] proposed a multimodal, fully convolutional network to extract meaningful semantic structures from document images. Based on a graph convolution based model, Liu *et al.* [109] combined textual and visual information presented in visually rich documents to perform entity recognition on document data. Zhang *et al.* [200] proposed a multimodal framework for simultaneous text reading and information extraction in visually rich documents for document understanding. By

utilizing the graphical property of business documents, Raja *et al.* [142] employed deep neural networks for table structure recognition. Madhav *et al.* [5] utilized an end-to-end trainable cascade deep architecture for table detection in document images. Olivier *et al.* [169] proposed a document retrieval model for answering questions on handwritten document image collections. Dauphinee *et al.* [38] constructed a model that uses both the visual information and the textual content of a given document to make a decision in a late fusion manner. Javier *et al.* [48] introduced an ensemble pipeline by combining image predictions with the text predictions produced by image and text modalities.

### Attention-based Representations

Attention learning was adopted to learn to attend to the most relevant regions of the input space in order to assign different weights to different regions. It was first proposed by Bahdanau *et al.* [15] for neural machine translation. The mechanism is firstly used for machine translation where the most relevant words for the output often occur at similar positions in the input sequence. Later, Vaswani *et al.* [172] proposed a self-attention module in machine translation models which could achieve state-of-the-art results at the moment. Then, the self-attention module was introduced to guide the visual attention from images. For the image modality, the self-attention-based modules learn to focus on particular image regions within a given document image [143, 179, 204]. Beyond the visual attention modules that are applied solely to the image modality, recent studies have introduced co-attention models that learn simultaneously from visual and textual attention to benefit from fine-grained representations of both modalities [79, 125]. Wang *et al.* [180] proposed a novel position-focused attention network to investigate the relation between the visual and textual views. Chen *et al.* [29] proposed a question-guided attention map that projects the question embeddings to the visual space, and formulates a configurable convolutional kernel to search the image attention region. Furthermore, some existing works that handled the task of jointly learning the interaction between image and text features used co-attention and self-attention modules [115, 192–194].

However, with such approaches the learning processes of the vision-language modalities are still independent one from another, and lack focusing on the inner relations and

the interactions between language and vision modalities. Therefore, some other works intended to exploit pre-training techniques for language-vision representation learning to construct a better multimodal representation space [59, 137]. These techniques have been exploited lately in document understanding tasks to learn more generic cross-modality representations between visual-textual information incorporated within documents. Aiming to alleviate the heterogeneity gap within and across modalities have shown that, when pre-trained in an end-to-end fashion on large amounts of data, these models learn more generic representations, and thus, yield to accurate performance when transferred to downstream tasks with low-scale datasets.

### Coordinated Representations

Multimodal document pre-training has seen increased attention recently as it allows to train semantically meaningful embeddings as a prior to a learnable downstream task. Given its great success from a NLP perspective, Devlin *et al.* [41] introduced a model called BERT, a deep bidirectional encoder based-transformers, which learns representations from unlabeled text by jointly conditioning on both left and right context. From a CV perspective, given the success of pre-training methods in NLP, Dosovitskiy *et al.* [46] extended the transformer [172] framework to introduce a transformer-based architecture applied directly to sequences of image patches to extract generic visual representations. Besides, the mechanisms used to leverage features from document modalities differ one from another. LayoutLMv1 [185] jointly models interactions between text and layout information across document images by adding 2D word positions in the language representation to better align the layout information with the semantic representation. LayoutLMv2 [184] leverages vision, language, and layout modalities in a cross-modal pre-training scheme for a better cross-modality interaction. In LayoutLMv3 [72], the authors propose a joint multimodal approach to model the interaction between textual, visual, and layout information in a unified multimodal pre-training network, with different pre-text tasks for a better generality to image-centric and text-centric downstream document AI tasks. SelfDoc [104] exploits cross-modal learning in the pre-training stage to perform a task-agnostic framework to model information across textual, visual, and layout infor-

mation modalities without requiring document data annotation. In DocFormer [9], the authors encourage multimodal interaction using a multimodal transformer architecture to perform visual document understanding. TILT [139] used bounding boxes of the OCRed words to serve as a region proposals, and add the region features to the corresponding language embeddings. UDOC [60] used document object proposals and concatenate Faster R-CNN region features with their language embeddings.

A broad category of pre-training techniques are those that use contrastive losses, which have been used in a wide range of CV applications like image-text similarity, and cross-modal retrieval [195, 198]. Such methods aim at mapping text and images into a common space, where semantic similarity across different modalities can be learned by ranking-based contrastive losses [59, 103, 114]. While dealing with vision-language sample pairs, though individual samples may demonstrate inherent heterogeneity in their content, they are usually coupled with each other based on some higher-level concepts such as their categories. This shared information can be useful in measuring semantics of samples across modalities in a relative manner. Verma *et al.* [173] analyzed the degree of specificity in the semantic content of a sample in the vision modality with respect to semantically similar samples in the language modality. Krishnan *et al.* [86] measured the similarity score between the word distributions across two document images, by detecting patterns of text re-usages across documents written by different individuals irrespective of the minor variations in word forms, word ordering, layout or paraphrasing of the content.

In the next section, we describe the methodology followed through our work in more detail, as well as our objectives and principal contributions.

## 1.5 Downstream Applications

This thesis comprises the application of three different kinds of downstream tasks. This includes well-established tasks in document understanding literature like document classification, and also two novel downstream tasks introduced in this thesis: few-shot document classification and content-based document retrieval. To perform these downstream applications, we make use of two publicly available benchmark document datasets, containing

samples of images from scanned documents from USA Tobacco companies, published by Legacy Tobacco Industry Documents and created by the University of California San Francisco (UCSF).

### 1.5.1 Datasets

Table 1.1: The Distribution of document pages over the RVL-CDIP and Tobacco-3482 datasets.

| | RVL-CDIP | | | Tobacco-3482 |
|---|---|---|---|---|
| Categories | #Training Data | #Validation Data | #Test Data | #Available Data |
| advertisement | 19,963 | 2,522 | 2,515 | 238 |
| budget | 20,010 | 2,485 | 2,505 | - |
| email | 19,954 | 2,530 | 2,516 | 611 |
| file folder | 20,012 | 2,451 | 2,527 | - |
| form | 19,957 | 2,537 | 2,506 | 441 |
| handwritten | 20,031 | 2,434 | 2,532 | - |
| invoice | 19,944 | 2,576 | 2,477 | - |
| letter | 20,103 | 2,430 | 2,464 | 580 |
| memo | 19,975 | 2,533 | 2,489 | 631 |
| news article | 19,987 | 2,526 | 2,463 | 190 |
| presentation | 20,043 | 2,468 | 2,489 | - |
| questionnaire | 20,042 | 2,516 | 2,435 | - |
| resume | 20,006 | 2,424 | 2,536 | 122 |
| scientific publication | 19,829 | 2,524 | 2,569 | 265 |
| scientific report | 19,984 | 2,508 | 2,498 | 271 |
| specification | 19,997 | 2,531 | 2,472 | - |
| note | - | - | - | 204 |

#### RVL-CDIP Dataset

The RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset is a subset of the IIT-CDIP Test Collection presented in [61]. This dataset consists of gray-scale labeled scanned document images into 16 classes (advertisement, budget, email, file folder, form, handwritten, invoice, letter, memo, news article, presentation, questionnaire, resume, scientific publication, scientific report, specification). The dataset is split into 320K training documents, 40K documents documents for validation and test sets. For notation simplicity, we denote the dataset as RVL-CDIP. Some representative images from the dataset are shown in Figure 1.11.

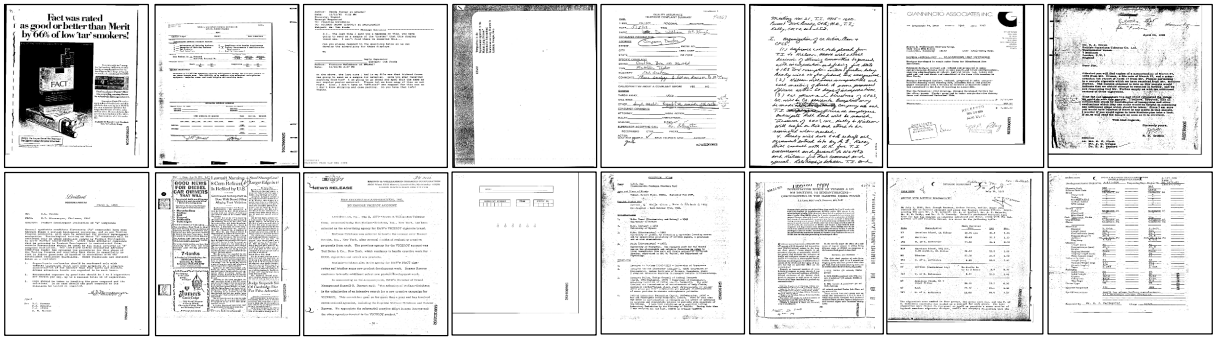


Figure 1.11: Samples of different document classes in the RVL-CDIP dataset which illustrate the low inter-class discrimination and high intra-class structural variations of document images. From left to right: *Advertisement, Budget, Email, File folder, Form, Handwritten, Invoice, Letter, Memo, News article, Presentation, Questionnaire, Resume, Scientific publication, Scientific report, Specification.*

#### Tobacco-3482 Dataset

The Tobacco-3482 dataset is a smaller sample containing 3,482 gray-scale document images presented in [90]. This dataset is formed by documents belonging to 10 classes not uniformly distributed, which are: ADVE, Email, Form, Letter, Memo, News, Notes, Report, Resume and Scientific. Some representative images from the dataset are shown in Figure 1.12.

### 1.5.2 Document Classification

The document image classification task aims to predict the category of visually rich document images. It is considered as one of the branches of scanned document image and text classification, where the classifier is able to tag a suitable class to the document from a list of predefined classes [3, 36, 61]. This makes the process of organizing and maintaining documents/data easy and efficient. Figure 1.13 presents an overview of the process of classifying document images based on two-stream deep neural networks.

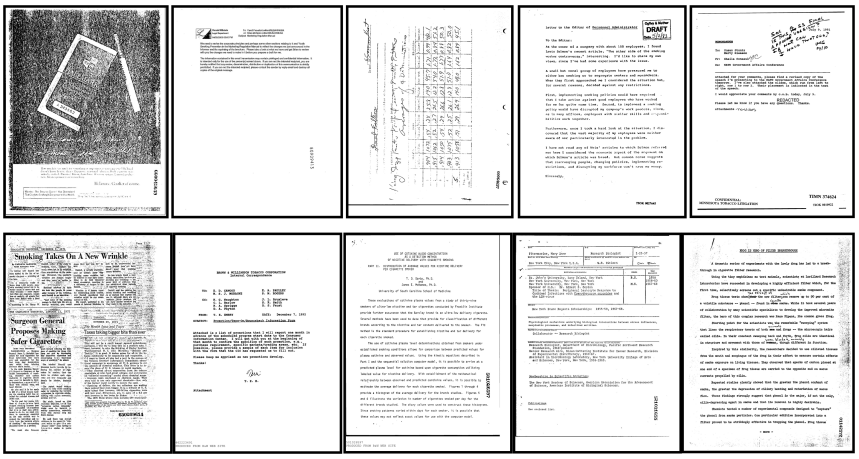


Figure 1.12: Samples of different document classes in the Tobacco-3482 dataset which illustrate the low inter-class discrimination and high intra-class structural variations of document images. From left to right: *ADVE, Email, Form, Letter, Memo, News, Note, Report, Resume, Scientific.*

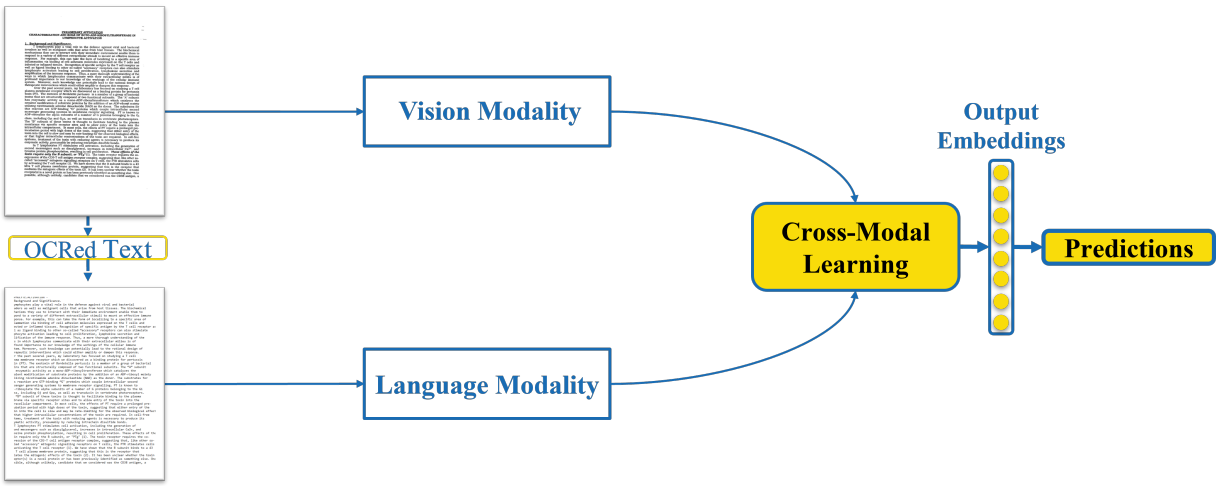



Figure 1.13: Overview of a multimodal deep neural network to perform cross-modal document image classification. The network is based on vision and language modalities.

### 1.5.3 Content-based Document Retrieval

Multimodal content-based document retrieval aims to identify relevant data across different modalities. The principal approach to address this task is to learn a joint semantic embedding space that can capture the inherent relationships between both modalities (see Figure 1.14). We aim to retrieve the category of the retrieved samples based on the

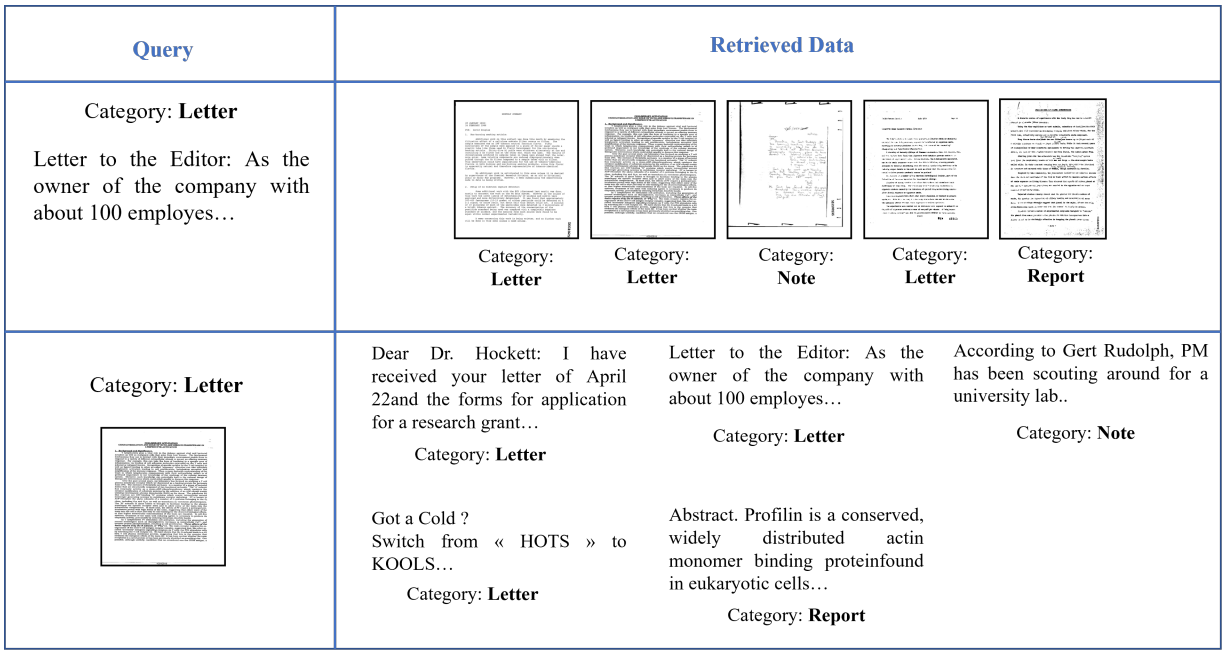



Figure 1.14: Examples of content-based document retrieval. The first query is given from the language modality. the expected results contain relevant and semantic visual representations. Then we retrieve the category of each result as Top-k retrieved samples which belong to the same category as the language query. The second query corresponds to the query document image from the vision modality. The goal is to retrieve relevant semantic information related to the query document image. Then, the category of each result is retrieved as Top-k retrieved samples belonging to the same category as the vision query.

given query sample. Specifically, retrieval involves computing the Euclidean distance between a query descriptor and every descriptor of the training set. The sorted distances are then used to rank the document images of the training data, and return a sorted list of documents. In the cross-modal content-based document retrieval context, given a query document image, we aim to retrieve meaningful semantic information related to the query, and then retrieve the category of each top-k ranked retrieved samples. The task of cross-modal retrieval has been a hot research topic in both computer vision and NLP communities. This is mainly carried on between images and text [177, 206]. The principal approach to address this task is to learn a joint semantic embedding space that can capture the inherent relationships between both modalities [175, 181].

### 1.5.4 Few-Shot Document Classification

Few-shot learning is a challenging problem as it has only limited data for training and needs to verify the performance on the data for unseen classes. An effective solution for few-shot classification problem is to apply a meta-learning (also called learning-to-learn with multi-auxiliary tasks) scheme on top of a pre-trained embedding network (see Figure 1.15). The key is how to robustly accelerate the learning progress of the network without suffering from over-fitting with limited training data [49, 141, 163, 174].

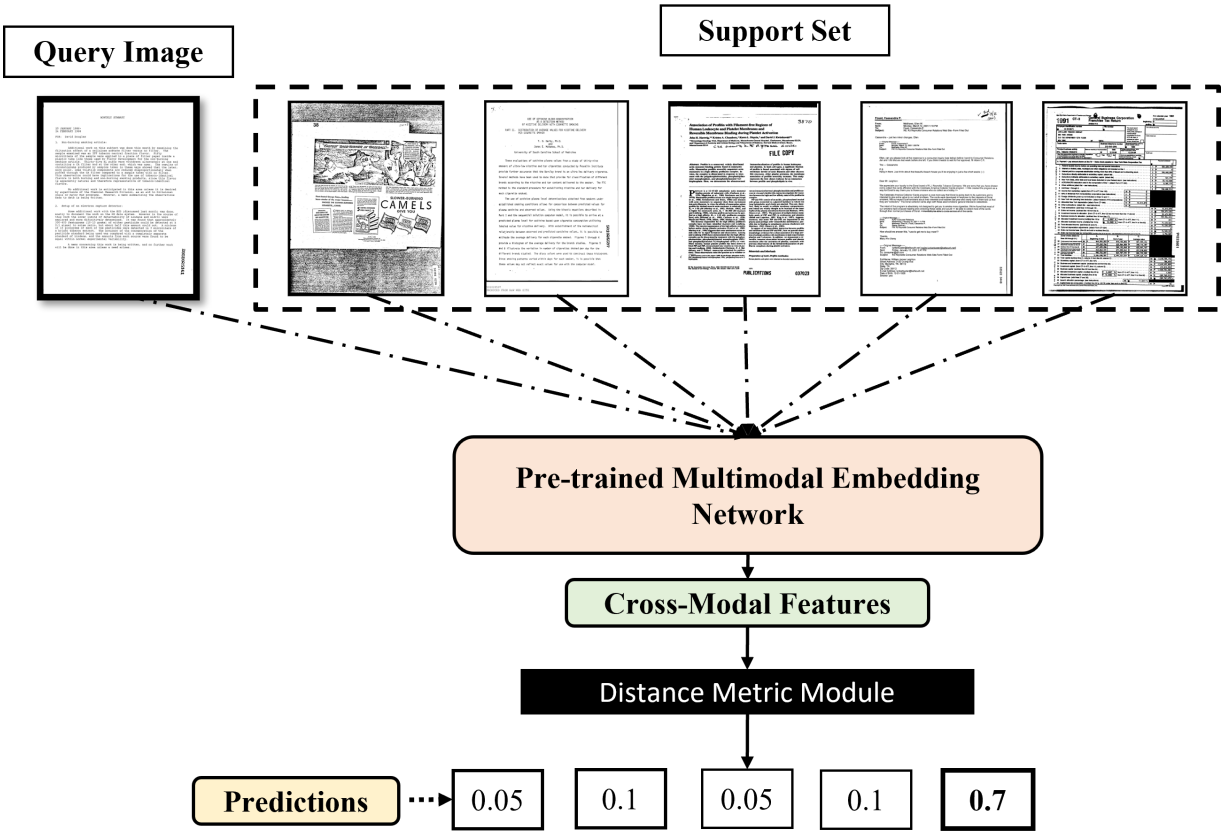



Figure 1.15: Meta-learning with an episodic task(5-way, 1-shot example). For each task, the training samples from the support set and the query samples are encoded by the embedding network. Query sample embeddings are compared with the centroid of training sample embeddings and make a further prediction.

In the next section, we describe the methodology followed through our work in more detail, along with our objectives and principal contributions.

## 1.6 Major Contributions

As we explained in the previous sections, multimodal document learning comprises a large set of challenges and applications. Although developments in this area have achieved outstanding performance in different applications such as document classification, named entity recognition, etc. Research in this field continually grows as improvements in the precision of these systems are demanded. In this dissertation, we propose to tackle the problem of multimodal document understanding through language and vision being the two principal data types. Our principal objective is to develop strategies that find a common semantic space that produces effective multimodal representations. We also aim to develop approaches easily adapted and evaluated for the downstream applications described earlier. The advantage of finding a common semantic space is to allow easily perform comparisons between target textual and visual content by mapping each modality to this space. This approach has been a successful strategy not only when working with vision and language, but also in the combination of multiple modalities. In this section, we present our contributions to the field of multimodal document understanding as well as a description of the methodology adopted in each one. Next, we present the organization of the document.

- **Chapter 2 - Multimodal Deep Feature Fusion.** In this chapter, we propose hybrid cross-modal deep networks based on deep learning techniques that leverage textual and visual data into a joint representation space. This objective seeks to achieve the development of systems that explore the semantic relationships between document images and their corresponding textual content that are easily adaptable to perform document classification. With this approach, we show that merging the two modalities with different fusion schemes enables the system to learn effective multimodal representations, and thus, boost the performance compared to single-modal networks. Moreover, we show that, dynamic word embeddings learn relevant semantic information from the text corpus compared to static word embeddings, as well as the ability of heavyweight deep neural networks to learn higher level features comparing to lightweight architectures. The proposed frameworks can handle any

given document image with its corresponding language content, projects them into a common space based on a feature fusion methodology, and sort out accurate predictions regarding the category of the given document. This goal is linked to our first two contributions that carry out the tasks of multimodal document classification in which the proposed frameworks have similarities. Our works present a new baseline on two benchmark document datasets. The results are presented in two published articles titled *"Cross-modal deep networks for document image classification"* [16] and *"Visual and textual deep feature fusion for document image classification"* [17].

- **Chapter 3 - Multimodal Deep Mutual Learning.** With this chapter, we explore and develop novel learning strategies that evaluate the impact of the quality of the data in the model performance when the problems of noisy text (*i.e.* where there is a lack of semantic meaning) are encountered. For example, some types of documents are mainly not recognizable by OCR algorithms, leading most of the time to losing textual information and semantic meaning. Thus, the visual information within the visual regions of the document should be strongly emphasized. Meanwhile, some other types of documents do not contain any visual spatial information, in which case a stronger emphasis on the textual information within the language cues is highly required. Understanding and analyzing document data properly enables us to create strategies able to leverage it aiming for good performance. This goal is linked to our third contribution where we present a mutual learning strategy to model the interaction between visual and textual features learned across the vision and language modalities throughout the learning stage. The mutual learning strategy encourages collaborative learning, allowing the vision and language modalities to simultaneously learn their discriminant features in a mutual learning manner. The main objective of this contribution is to enable our framework to be efficient in improving not only the overall performance of the multimodal fusion modality, but also the performance of the single-modal modalities. The results of this work are presented in the journal article *"EAML: ensemble self-attention-based mutual learning network for document image classification"* [18]

- **Chapter 4 - Multimodal Document Representation Learning.** The interpretation of a piece of content in document data relies heavily on its semantic meaning. For example, a heading can indicate and summarize the meaning of subsequent blocks of text, and a text sequence could be useful for understanding the type of the document. In contrast to other data formats like images or plain text, documents combine textual and visual information, and both of the two modalities are complemented by the document layout. From a practical perspective, many tasks related to document understanding are label-scarce. A framework that can learn from unlabeled documents (*i.e.* pre-training) and perform model fine-tuning for specific downstream applications is more preferred than the one that requires fully-annotated training data. This goal is linked to our fourth contribution that carries out the downstream tasks of multimodal document classification, multimodal content-based retrieval, and few-shot document classification. This chapter presents our fourth contribution. We design a unified network for cross-modal representation learning. Our network consists of leveraging two flexible extra levels of cross-modal interactions through co-attention module, to capture high-level interactions between vision-language cues in document images. The proposed approach shows its superiority over the uni-modal methods. A superior performance shows that a good generalization has been achieved which enables to classify the documents in different domains. The results of this work will be presented in the journal article titled *"VL-CDoC: Vision-Language Contrastive Pre-Training Model for Cross-Modal Document Classification"* [19], which is currently in the process of revision and submission.

- **Chapter 5 - Improved Multimodal Semantic Representation Learning.** With this last contribution, we intend to decrease the gap between vision-language models and vision-language-layout prior works and extend our last framework from Chapter 4 to perform fine-tuning on three downstream applications: multimodal document classification, multimodal content-based document retrieval and few-shot document classification, when insufficient labelled data are present. The objective of this chapter is to encourage multimodal interaction from language and vision in a self-supervised learning manner. We propose a framework that enables to pre-

train multimodal transformers with a two-step approach where feature learning and clustering are decoupled. Our network is first pre-trained with a nearest-neighbour instance discrimination technique to obtain semantically meaningful features. Then, the obtained features are used as a prior in a learnable clustering approach to remove the ability for cluster learning to depend on low-level features. The introduced framework has shown its effectiveness on three main downstream applications which are: document classification, few-shot document classification, and content-based document retrieval. The results of this work will be submitted in a conference article titled *"LSRD: Learning Improved Semantic Representations for Document Understanding"*.

This dissertation is organized as follows. Chapter 2 presents our proposed approaches related to cross-modal feature fusion learning to perform document classification, along with the experimental settings, and ablation studies performed to demonstrate the effectiveness of the proposed approaches in two studies: first, the different feature fusion methodologies to leverage visual-textual features into a common representation space, and second, an extended evaluation of the impact of training static and dynamic word embeddings, the heavyweight and lightweight DCNNs on the classification performance of document data. Chapter 3 addresses the limitations of Chapter 2 and presents a collaborative mutual learning strategy to transfer positive information from one modality to another, enabling to improve the accuracy results for each category, demonstrated in the ablation studies. Chapter 4 presents details about document pre-training aiming for a better multimodal document understanding. This chapter comprises the description of the experimental pretraining settings, as well as the results obtained on two downstream applications: document classification, and cross-domain few-shot learning. Chapter 5 presents a more general and model-agnostic pre-trained model for document understanding applications which are: document classification, content-based document retrieval, and few-shot learning. Finally, Chapter 6 presents general conclusions of the proposed developments during this thesis along with the future ideas for the future research.

## 1.7 List of Publications

The publications originating from this thesis are as follows:

### 1.7.1 International Journals

J1 : **IJDAR. Bakkali Souhail**, Ming Zuheng, Coustaty Mickael and Rusiñol Marçal (2021). EAML: ensemble self-attention-based mutual learning network for document image classification. *International Journal on Document Analysis and Recognition (IJDAR), 24(3), 251-268.*

J2 : **PR (Under Review). Bakkali Souhail**, Ming Zuheng, Coustaty Mickael, Rusiñol Marçal and Terrades Oriol Ramos (2022). VLCDoC: Vision-Language Contrastive Pre-Training Model for Cross-Modal Document Classification. *arXiv preprint arXiv:2205.12029.*

### 1.7.2 International Conferences

C1 : **ICIP. Bakkali Souhail**, Ming Zuheng, Coustaty Mickael and Rusiñol Marçal (2020, October). Cross-modal deep networks for document image classification. *In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 2556-2560). IEEE.*

### 1.7.3 International Workshops

W1 : **CVPRW. Bakkali Souhail**, Ming Zuheng, Coustaty Mickael and Rusiñol Marçal (2020). Visual and textual deep feature fusion for document image classification. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 562-563).*

# CHAPTER 2

# Multimodal Deep Feature Fusion

*Fusion will be the final way out for the future.*

*– Shen Wenquan*

## 2.1 Motivation

In our first approximation to the topic of document image classification -which has been explored extensively over the past few years- we adopt a two-stream neural architecture for cross-modal feature fusion. Most recent approaches handled this task by jointly learning the visual features of document images and their corresponding textual content. Due to the various structures of document images, the extraction of semantic information from its textual content is beneficial for document image processing tasks. Given their natural design, we aim to solve the following research question of **how to develop and efficiently implement deep network-based models for discriminative and compact cross-modal representations.**

In this chapter, we conduct an exhaustive investigation of nowadays widely used deep networks as well as word embedding procedures used as the main backbones in our two-

stream network, in order to extract both visual and textual features from document images. Moreover, a joint feature learning approach that combines visual features and textual embeddings is introduced as an early fusion methodology. Our goal is to evaluate the representation learning capability of our two-stream deep network to encode meaningful information from the vision and language modalities to perform the cross-modal document classification task.

Recent advances in deep learning techniques have made significant progress in many areas in CV and NLP. The main reason for such success is the ability to train a deep learning model that can retain profound knowledge from large-scale labeled dataset such as the ImageNet dataset [87] used for image classification. From a computer vision perspective, the concept of transfer learning from the object recognition domain was used to improve the recognition accuracy on smaller datasets [150]. To investigate this approach more efficiently, we train our vision modality using ImageNet weights as it has shown to be effective in earlier attempts on document image classification [4]. Several research studies in the literature have been using deep neural networks for document analysis tasks. They focused on the structural similarity constraints and the visual features of document images [26, 89, 90, 156]. As most recent deep learning methods do not require extracting features manually, the state-of-the-art approaches based on visual information of document images treated the problem as a conventional image classification task. Additionally, from a natural language processing perspective, Yang *et al.* [189] presented a neural network to extract meaningful semantic information based on word embeddings from pre-trained natural language models. Nevertheless, classifying documents with only visual information may encounter the problem of low inter-class discrimination, and high intra-class structural variations of highly overlapped document images [3]. As such, jointly learning visual cues and text semantic relationships is an inevitable step to mitigate the issue of highly correlated classes. Recent methods have used multimodal techniques to leverage both vision and language modalities extracted by an optical character recognition OCR engine [73] to perform fine-grained document image classification [11, 12, 38, 185].

Therefore, we study the capability of static and dynamic word embeddings to extract meaningful information from the text corpus. While static word embeddings fail to capture

polysemy, by generating the same embedding for the same word in different contexts, dynamic word embeddings are able to capture word semantics in different contexts to address the issue of polysemous and context-dependent nature of words. We explored and evaluated both static and dynamic word embeddings on the large-scale RVL-CDIP [1] [61] dataset. Furthermore, we propose in this chapter a two-stream cross-modal deep neural network to learn simultaneously from the visual structural properties and the textual information from document images based on two different models. The learnt cross-modal features are combined as the final representation of our proposed network to boost the classification accuracy of document images. However, to perform text classification, an OCR is employed to extract the textual content of each document image, followed by a latent semantic analysis. We utilize the pre-trained Glove and FastText [119, 135] models as two static word embeddings, followed by a gated recurrent unit (GRU) mechanism introduced by J.Chung *et al.* . and K.Cho *et al.* [33]. GRU is a simplified variant of LSTM architectures introduced by S. Hochreiter and J. Schmidhuber [58] to overcome the vanishing gradient problems. Moreover, based on both left and right context, the deep bidirectional pre-trained $\text{Bert}_{Base}$ model [41] is utilized as a contextualized dynamic word embedding to learn the textual semantic features.

To conduct the document image classification task, we investigate the impact of both heavyweight (*i.e.* with a large amount of parameters) and lightweight (*i.e.* with a much lower number of parameters) deep network architectures on learning deep structural properties from document images. These models have been chosen for their performance on the ImageNet [39] dataset at different levels of computational and time cost, starting from models operating in a constrained computational environment for mobile applications (*i.e.* $\text{NasNet}_{Mobile}$ [209]), to computationally-heavy models (*i.e.* Inception-ResNet-v2 [164], $\text{NasNet}_{Large}$ [209]) designed to achieve real-time accurate results. The heavyweight models with large size parameters such as $\text{NasNet}_{Large}$, and Inception-ResNet-v2 can achieve state-of-the-art classification accuracy on the widely used ImageNet [39] dataset in the cost of the computational complexity and time consuming. Instead, the lightweight models with fewer parameters designed for the constrained environment (*e.g.* real-time environment),

[1] https://www.cs.cmu.edu/~aharley/rvl-cdip/

for mobile applications with less hardware resources, focus on the trade-off between the efficiency and the model accuracy.

The analysis of document data present in both document datasets, we found that amongst all classes, some samples from specific categories present particular layout properties and document structures as illustrated in the Figure 2.1. Most classes are mainly composed of text information such as Report, while the classes like Advertisement, and File Folder contain only images with very little text information. Specifically, some samples do not contain any text data. Another class such as Handwritten, which is composed of handwritten text characters, produces noisy output text resulted by the processing of the OCR engine. The idea behind this chapter relies on whether combining the learnt visual features with the learnt textual features could be effective in enhancing the feature representation space, and thus, achieving better yet effective results for the specific categories mentioned above (*i.e.* Advertisement, File Folder, Handwritten).

In summary, the main contributions of this chapter are as follows:

- We propose a two-stream cross-modal deep network that leverages both the learned textual embeddings and visual features to classify document images. We show that the proposed joint learning methodology boosts the overall accuracy compared to the single-modal networks.
- We introduce two feature fusion methodologies to merge vision-language features in the cross-modal framework.
- We evaluate the performance of static and contextualized dynamic word embeddings to classify textual content of document images.
- As well, we review the impact of training heavyweight and lightweight deep neural networks on learning relevant structural information from document images.

## 2.2 Approach

This section briefly presents the deep convolutional neural networks and word embedding procedures used in this chapter. On the one hand, we intend to investigate the impact of

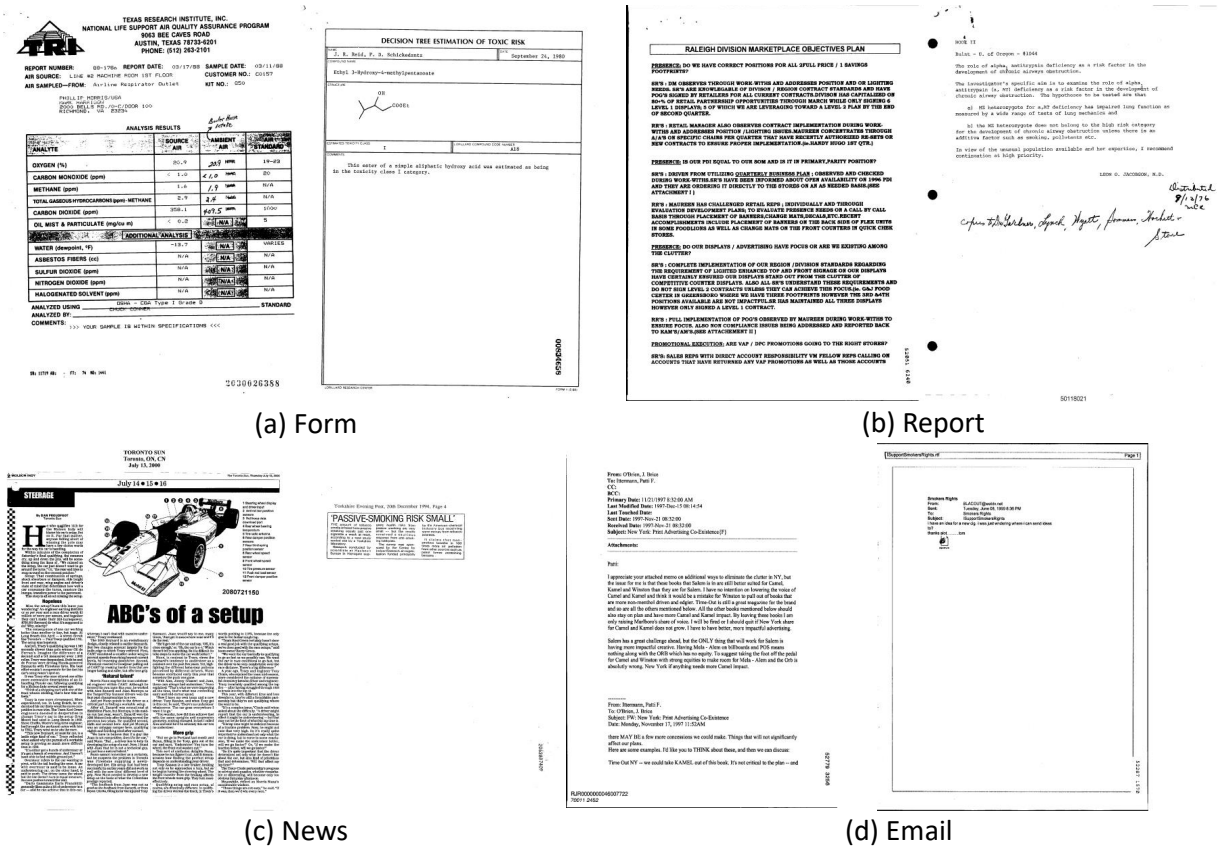



Figure 2.1: Sample images from the benchmark Tobacco-3482 dataset showing the low inter-class and high intra-class of structural variations of document images.

training lightweight and heavyweight deep networks on the classification performance on the RVL-CDIP and Tobacco-3482 datasets. On the other hand, we attempt to compare the performance of static and dynamic word embedding procedures used to generate features to process the text classification task. Figure 2.2 illustrates the proposed cross-modal network with $\text{NasNet}_{Large}$ and $\text{Bert}_{Base}$ as the vision and language backbones respectively.

### 2.2.1 Vision Modality

For the document visual embeddings, we propose to explore two well-known deep CNNs (NasNet and Inception-ResNet-v2) as main backbones to extract the image features.

**NasNet-A**(6@4032): The NasNet architecture [209] is composed of two types of layers: Normal layer, and Reduction layer. The Normal layer is a convolutional layer that returns a feature map of the same dimension, where the Reduction layer is a convolutional layer

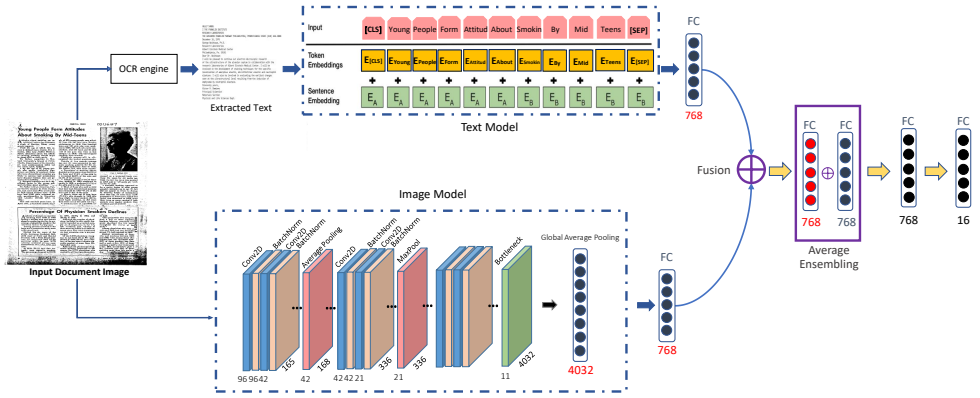



Figure 2.2: The proposed cross-modal deep neural network. The NasNet$_{Large}$ model is used for the vision modality, while Bert$_{Base}$ model is used for the language modality[16, 17]

.

that returns a feature map, where the feature map height and width is reduced by a factor of two. For NasNet-A(6@4032), 6 means $N = 6$, *i.e.* number of layers repeated, $4,032$ means the number of filters in the penultimate layer of the network. It has $88.02M$ parameters. We denote the model as NasNet$_{Large}$.

**NasNet-A**(4@1056): A second architecture based on the same network was studied with $N = 4$ layers repeated and $1,056$ filters in the penultimate layer of the network. This light network only has $4.23M$ parameters. We denote it as NasNet$_{Mobile}$.

**Inception-ResNet-v2**: The Inception-ResNet-v2 [164] architecture is a convolutional neural network that achieved state-of-the-art results on the ILSVRC image classification benchmark. Inception-ResNet-v2 is a variation of the earlier Inception-V3 model by introducing the bypass connection as in ResNet [63]. The model has $54.36M$ parameters.

### 2.2.2 Language Modality

For the textual part of documents, we use three well-known word-embeddings mixing static and dynamic approaches to perform text classification.

**GloVe**: GloVe [135] is an unsupervised learning algorithm that generates word embeddings by aggregating global word-word co-occurrence matrix from a corpus. The resulting embeddings show interesting linear substructures of the words in a vector space. We use

the pre-trained GloVe model on Wikipedia 2014 with Gigaword 5 ($6B$ tokens, $400K$ vocab, uncased, $50d$ vectors) parameters.

**FastText**: FastText [119] is a library for efficient learning of word representations and sentence classification. FastText breaks words into several character n-grams, which allows computation of word representations for words that did not appear in the training data, known as out-of-vocabulary words. We use the pre-trained FastText model on 2 million word vectors trained on Common Crawl ($600B$ tokens), and uses $1,999,996$ word vectors.

**Bert**: Bert [41] is a contextualized bi-directional word embedding based on the transformer architecture. Bert representations are jointly conditioned on both left and right context in all layers, using a faster highly-efficient attention-based approach. The $\text{Bert}_{Base}$ model we use consists of 12 attention layers, 768 hidden layers, 12 heads, $109M$ parameters, and uses a vocabulary of $30,522$ words.

The next section presents in detail the components of each modality of our proposed cross-modal deep neural network.

## 2.3 Cross-Modal Feature Learning

In this section, we present in detail the proposed cross-modal deep neural network for document image classification. In the first stream, we feed input document images to the vision backbone. In the second stream, we extract the textual corpus from document images with an OCR engine. Then, we feed the text corpus generated as the input to the word embedding backbone. Finally, we consider an early fusion process to merge the two modalities to enhance the performance of the cross-modal modality compared to the single-modalities.

### 2.3.1 Visual Features

Deep CNNs have exhibited their exceptional performance in both general image recognition and image classification tasks. Since transfer learning has shown its effectiveness while transferring to smaller datasets, we train the three deep CNNs discussed above using the pre-trained ImageNet weights. The vision modality extracts visual features that

are passed to a global average pooling layer to reduce the spatial dimensions of a three-dimensional tensor. It performs also a more extreme type of dimensionality reduction. For the final layers of the three deep CNNs, the global average pooling layer is passed to the last fully connected layer to perform classification with a softmax layer. The categorical cross-entropy loss function of softmax is given by:

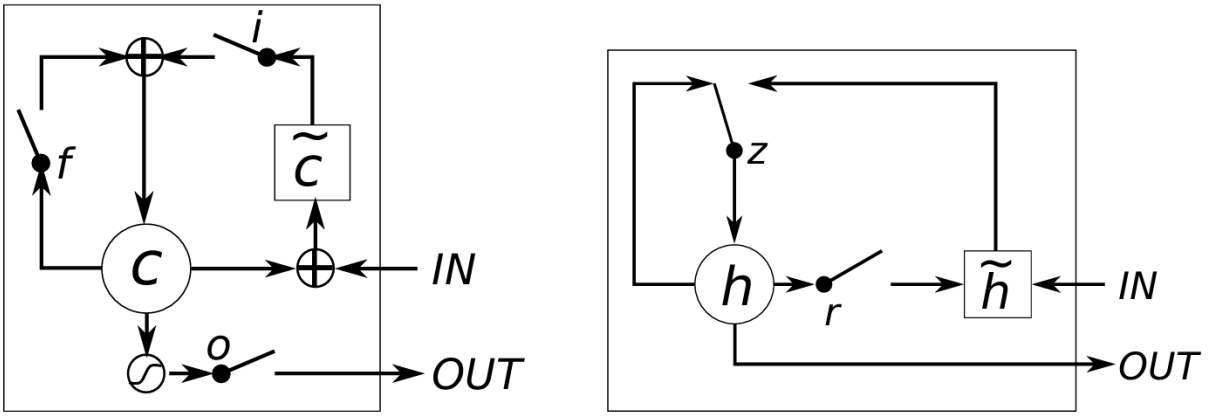
$$\begin{aligned}\mathcal{L}_{s1}(\mathbf{X}_1;\Theta_1) &= \sum_{k=1}^{K} -y_k log P(\hat{y}_k|\mathbf{X_1},\theta_k) \\ &= -\sum_{k=1}^{K} y_k log \frac{e^{f^{\theta_k}(\mathbf{X}_1)}}{\sum_1^K e^{f^{\theta_{k'}}(\mathbf{X}_1)}}\end{aligned} \tag{2.1}$$

where $\{\mathbf{X}_1, \Theta_1\} \in \mathbb{R}^{d_1}$, and $d_1$ is the dimension of $X_1$ features of the vision modality. $K$ is the number of classes in the dataset where $K = 16$, $y_k$ is the one-shot label of the feature $\mathbf{X_1}$, $P(\hat{y}_k|\mathbf{X_1},\theta_k)$ is the estimated probability of $y_k$ calculated by the softmax function over the activation function $f^{\theta_k}(\mathbf{X_1})$, where $\{\theta_k\}_{k=1}^{K} = \Theta_1$, $\theta_k \in \mathbb{R}^{d_1}$. The bottleneck layer of the image branch is extracted as the feature $\mathbf{X}_1$ of the input image.

### 2.3.2 Textual Features

As textual content is required to perform text classification, we process all document images with an off-the shelf optical character recognition (OCR) engine, *i.e.* Tesseract OCR[2] [160]. It is based on LSTM layers and includes a neural network subsystem configured in English as a text line recognizer. Besides, the OCRed text extracted is noisy and not clean due to the different ways of presenting documents from plain, handwritten, and curved text, exotic fonts, multi-column layouts, the wide variety of tables, forms, and figures. Many word embeddings process a good tokenization of the words by getting the embedding (*i.e.* a vector of real numbers) for each word in the sequence, where each word is mapped to a *emb_dim* dimensional vector that the model will learn during training. In average, for GloVe word embedding, we found $3,581,896$ unique tokens and a total number of $400,000$ word vectors on the RVL-CDIP corpus. As well, we found $3,601,377$ unique tokens, $24,109$ of null word embeddings, and a dictionary size of $3,601,377$ for

[2] https://github.com/tesseract-ocr/tesseract




(a) Long Short-Term Memory (b) Gated Recurrent Unit

Figure 2.3: Illustration of (a) LSTM and (b) gated recurrent units (GRU). (a) i, f and o are the input, forget and output gates, respectively. c and $\tilde{c}$ denote the memory cell and the new memory cell content. (b) r and z are the reset and update gates, and h and $\tilde{h}$ are the activation and the candidate activation.

FastText word embedding on the same standard dataset. Contrary to traditional shallow representations (*i.e.* Word2Vec [56], GloVe [135], FastText [119]), as they fail to capture higher-level information, many different dynamic word embedding procedures (*i.e.* ELMO [137], Bert [41], XLNet [190]) have been proposed to capture semantic meaning to deal with the context-dependent nature of words. For the $\text{Bert}_{Base}$ model, we processed the tokenization by splitting the input text into a 128 sequence list of tokens. To deal with out of vocabulary (OOV), $\text{Bert}_{Base}$ uses a WordPiece tokenization technique in which every OOV word is split into sub-words. The input embeddings are then computed by summing the corresponding word embeddings, and segment embeddings. Then, the input embeddings are passed to the attention-based bidirectional transformer. After preprocessing the textual content extracted by the OCR engine from document images, we pass the input embeddings of both Glove and FastText to a GRU network of 32 nodes and 3 hidden layers. Figure 2.3 illustrates the LSTM and GRU units taken from [33]. The final layers of the three models are passed to a softmax layer with categorical cross-entropy loss function.

### 2.3.3 Cross-Modal Features

In this part, we intend to study the effectiveness of the cross-modal features that are jointly learned from the vision and language modalities to classify document images. We adopt an early fusion process with two different methodologies, (*i.e.* equal concatenation, and average ensemble fusion). We assume that the dimension of the features extracted from the vision modality or the language modality is denoted as $d$.

**(a) Equal Concatenation:** We add a fully connected layer to the vision modality, having the same dimensional output vector as the language modality. The final cross-modal features are the concatenation of the two equal embedding features given by:

$$X_a = [X_1|X_2]; \quad X_a \in \mathbb{R}^{2d_1} \tag{2.2}$$

where $X_1 \in \mathbb{R}^{d_1}$ is the obtained image embedding feature, and $X_2 \in \mathbb{R}^{d_2}$ is the text embedding feature, $d_1 = d_2$ and $|$ is the concatenation operation.

**(b) Superposing Fusion:** We employ a pixel-wise addition between the image and text embedding features, (*i.e.* superposing directly the two embeddings to generate the cross-modal features). Note that the obtained cross-modal features have the same dimension as the image or text embedding features.

$$X_{Av} = [X_1 + X_2]; \quad X_{Av} \in \mathbb{R}^{d_1} \tag{2.3}$$

#### Training Protocol

The learning of the cross-modal features include two main parts: the learning of the parameters of the vision modality $\Theta_1$ and the parameters of the language modality $\Theta_2$. Then, the parameters of the network $\Theta = \{\Theta_1, \Theta_2\}$ are optimized by the global cross-entropy loss function $\mathcal{L}(\Theta)$ given by:

$$\mathcal{L}(\Theta) = \sum_{k=1}^{K} -y_k log P(\hat{y}_k|\mathbf{X}, \Theta) \tag{2.4}$$

where $\mathbf{X}$ is the cross-modal features $\mathbf{X_a}$ or $\mathbf{X_{Av}}$.

## 2.4 Experiments and Analysis

To evaluate the performance of our proposed ensemble trainable network, we make use of the two benchmark datasets RVL-CDIP and Tobacco-3482 introduced in the Section 1.5.1.

### 2.4.1 Preprocessing

As the deep convolutional neural networks (DCNNs) used in this chapter require fixed size input images, we first downscale all document images presented in both RVL-CDIP and Tobacco-3482 datasets to the expected input size of the networks. The original document images size is about $1000x750$ pixels. For the NasNet$_{Large}$ backbone, document images are resized to $331x331$ pixels. For the Inception-ResNet-v2 backbone, the images are resized to $299x299$ pixels, and resized to $224 \times 224$ for NasNet$_{Mobile}$. As a data augmentation typical step, we intended to minimize the high intra-class similarity variations in document images. To do so we applied shear transform with a range of 0.1 as in [166]. This technique is a common practice to stochastically transform each input during stochastic gradient descent (SGD) training [7], to artificially enlarge the training data in order to improve the performance. Also, we randomly shifted images horizontally and vertically with a range of 0.1. For effective training, we introduced cutout data augmentation [42] that has shown its efficiency towards improving regularization of DCNNs. It consists of randomly masking a square region in an image at every training step, thus removing the redundancy of the images and augmenting the dataset by partially occluded versions of existing samples. As a final pre-processing step for vision modalities, we convert the gray-scaled document images to RGB images.

Intuitively, the text corpus fed to the input layer of the text branch was extracted with an off-the-shelf optical character recognition OCR (*i.e.* Tesseract OCR). We utilized this OCR engine to conduct a fully automatic page segmentation, as the document images from the datasets are well-oriented and relatively clean. Hence, we run the Tesseract OCR engine. We used the version $4.0.0 - beta.1$ of Tesseract based on a LSTM engine to aim for better accuracy. Also, a fully automatic page segmentation without orientation or script detection is conducted. The resulting extracted text was not post-processed.

Although document information might be lost in OCR, such as typeface, graphics, layout, stop words, mis-spellings, symbols and characters. It could benefit from some level of spell checking to improve the semantic learning. However, we chose to provide the true output of Tesseract OCR as it is.

### 2.4.2 Implementation Details

In this subsection, we describe the implementation details used to train the proposed single-modal and cross-modal approaches. We have trained all networks on a NVIDIA Quadro GP100 GPU, using stochastic gradient descent optimizer (SGD), with a momentum of 0.9, a learning rate of $1e-3$, and a step decay schedule defined as:

$$\mathrm{Lr} = \mathrm{initial\_lr} * drop^{\left(\frac{\mathrm{iter}}{\mathrm{iter\_drop}}\right)} \tag{2.5}$$

where drop and iter_drop took values of 0.5.

The visual modalities were trained with a batch size of 16 for 50 epochs. Early stopping was considered within 5 epochs to stop training once the performance of the model stops improving on the hold out validation dataset. Further, $L_2$ regularization was adopted to add a penalty for weight size to the loss function. Dropout was also applied to the final softmax layer with a probability of 0.5. For the language modality, it was trained with a batch size of 40, and a sequence length of 128 for 50 epochs. The cross-modal feature learning approach was fine-tuned using document pretraining weights obtained by the single modalities. We froze all layers except the last fully connected layers and trained our cross-modal network with both the equal concatenation and the superposing fusion methods, followed by the softmax layer to perform the final task of document image classification.

### 2.4.3 Overall Evaluation

#### Overall Evaluation on the Tobacco-3482 Dataset

Table 2.1: Overall accuracy on the Tobacco-3482 dataset versus model. E.C refers to Equal Concatenation, and S.F refers to Superposing Fusion.

| Model | Accuracy(%) | ADVE | Email | Form | Letter | Memo | News | Notes | Report | Resume | Scientific |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **single-Modal (Vision)** | 96.25 | 1 | 1 | 0.96 | 0.94 | 0.98 | 1 | 0.90 | 1 | 0.78 | 0.90 |
| **single-Modal (Language)** | 97.18 | 0.97 | 0.99 | 0.98 | 0.93 | 0.97 | 0.98 | 0.89 | 1 | 0.96 | 0.95 |
| Ensemble [12] | 87.8 | 0.93 | 0.98 | 0.88 | 0.86 | 0.90 | 0.90 | 0.85 | 0.71 | 0.96 | 0.68 |
| Two Stream Model [11] | 95.8 | 0.94 | 0.98 | 0.95 | 0.98 | 0.97 | 0.97 | 0.88 | 0.92 | 1 | 0.93 |
| **Cross-Modal** (E.C) | 98.42 | 0.98 | 0.99 | 0.95 | 1 | 0.98 | 0.97 | 1 | 1 | 0.96 | 0.98 |
| **Cross-Modal** (S.F) | 99.71 | 1 | 1 | 0.97 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

On the low-scale Tobacco-3482 dataset, the adopted cross-modal fusion methodologies achieve state-of-the-art performance. We report the overall accuracy results in Table 2.1, with the superposing fusion scheme achieving the best performance of 99.71% classification accuracy.

#### Overall Evaluation on the RVL-CDIP Dataset

On the large-scale RVL-CDIP dataset, all of the adopted networks in this work achieve comparable performance with the state-of-the-art results. We report the overall accuracy results in Table 2.2. The heavyweight $\text{NasNet}_{Large\ (768d)}$ model performs the best for our vision modalities at an accuracy of 91.45%, outperforming the other tested models $\text{NasNet}_{Large\ (4032d)}$, Inception-ResNet-v2, and $\text{NasNet}_{Mobile}$ at an accuracy of 91.12%, 85.04%, and 81.54% respectively.

As for the language modalities, the $\text{Bert}_{Base}$ model achieves comparable performance with the state-of-the-art results on the same benchmark dataset, with an accuracy of 84.96%. $\text{Bert}_{Base}$ manages to improve the performance thanks to its attention-based mechanism, while Glove and FastText still achieve good results on the text classification task at an accuracy of 71.54%, and 77.31% respectively. As each single modality is trained independently one from another, merging both modalities boosts the performance significantly for the two fusion modalities to 96.94%, 97.05% classification accuracy for equal

Table 2.2: The overall accuracy of the proposed methods with different backbones and different fusion modalities on the RVL-CDIP dataset. E.C refers to Equal Concatenation, and A.E refers to superposing Fusion.

| Method | Model | Acc.(%) | Top-5 Acc. | Precision | Recall | F1-Score | #Params |
|---|---|---|---|---|---|---|---|
| Baselines | Harley *et al.* [61] | 89.80 | - | - | - | - | - |
| | Nicolas *et al.* [12] | 90.06 | - | - | - | - | - |
| | Csurka *et al.* [35] | 90.70 | - | - | - | - | - |
| | Tensemeyer *et al.* [166] | 90.94 | - | - | - | - | - |
| | Afzal *et al.* [4] | 90.97 | - | - | - | - | - |
| | Das *et al.* [36] | 91.11 | - | - | - | - | - |
| | Das *et al.* [36] | 92.21 | - | - | - | - | - |
| | Dauphinee *et al.* [38] | 93.03 | - | - | - | - | - |
| | Dauphinee *et al.* [38] | 93.07 | - | - | - | - | - |
| | Xu *et al.* [185] | 94.42 | - | - | - | - | 160 M |
| Language-only | Glove-GRU | 71.54 | 93.86 | 0.75 | 0.72 | 0.72 | 179 M |
| | FastText-GRU | 77.31 | 95.15 | 0.80 | 0.78 | 0.78 | 30.47 M |
| | $\text{Bert}_{Base}$ | 84.96 | 96.74 | 0.86 | 0.86 | 0.85 | 109.19 M |
| Vision-only | $\text{NasNet}_{M}$ | 81.54 | 97.29 | 0.84 | 0.83 | 0.83 | 4.23 M |
| | Inception-ResNet-v2 | 85.04 | 97.80 | 0.88 | 0.86 | 0.87 | 54.36 M |
| | $\text{NasNet}_{L_{4032d}}$ | 91.12 | 98.61 | 0.92 | 0.91 | 0.92 | 84.98 M |
| | $\text{NasNet}_{L_{768d}}$ | 91.45 | 98.60 | 0.92 | 0.92 | 0.92 | 88.02 M |
| vision+Language | Cross-Modal (E.C) | **96.94** | **99.83** | **0.97** | **0.97** | **0.97** | 197.22 M |
| | Cross-Modal (A.E) | **97.05** | **99.85** | **0.97** | **0.97** | **0.97** | 197.21 M |

concatenation and superposing respectively. Thus, exceeding the current state-of-the-art results by a 2.63% margin.

### 2.4.4 Ablation Study

#### Evaluation on the Tobacco-3482 Dataset

To evaluate the effectiveness of our proposed cross-modal approach for document image classification, we firstly investigate the performance of the single modalities based on visual and textual features. Then, we compare our cross-modal method to the single modalities, and finally, to the state-of-the-art baselines based on two-stream deep neural networks.

In this part of evaluation, we propose to use $\text{NASNet}_{Large}$ to classify the document images with only visual features. As shown in Table 2.3, the $\text{NASNet}_{Large}$ gains the best result of 96.25% which outperforms the state-of-the-art single-modal method based on the InceptionV3 network by a 3.05% margin. Note that the $\text{NASNet}_{Large}$ is pre-trained on ImageNet, used as weight initialization as transfer learning is known to improve the

Table 2.3: Evaluation of the Vision modality against Baselines on Tobacco-3482 dataset.

| Method | Accuracy(%) |
|---|---|
| AlexNet [4] | 90.04 |
| GooGleNet [4] | 88.4 |
| VGG-16 [4] | 91.01 |
| ResNet-50 [4] | 91.13 |
| MobileNetV2 [12] | 84.50 |
| InceptionV3 [11] | 93.2 |
| **NASNet**$_{Large}$ | 96.25 |

Table 2.4: Accuracy comparison of Language-stream state-of-the-art models on the Tobacco-3482 dataset.

| Method | Accuracy(%) |
|---|---|
| FastText-CNN [12] | 73.8 |
| Feature Ranking (ACC2) [11] | 87.1 |
| Glove-CNN1D-LSTM | 51 |
| Glove-GRU | 61 |
| **Bert**$_{Base}$ | 97.18 |

classification performance significantly although the images are substantially different. Amongst all the current state-of-the-art baselines, we managed to push the performance much further by 3.15%.

Besides, for the single-modal language pipeline, we tested combined architectures such as CNN-LSTM and GRU on top of Glove word embeddings as shown in Table 2.4. Results demonstrate that the Bert$_{Base}$ model achieves a new state-of-the-art result of 97.18%, outperforming all existing methods with a very high margin of 10.08%. Therefore, attention-based approaches are highly-efficient operations thanks to their fast run-time characteristics.

In addition, Table 2.1. compares the performance of the two proposed fusion methods to perform cross-modal document image classification. For Equal Concatenation (E.C) feature fusion operation, we compress the visual features and concatenate them with the textual features, having both the same dimensional feature vector. As well, our cross-modal network manages to raise the performance for all classes except for the classes News and Form, where it drops by 1%. This is mainly due to the highly overlapped

Table 2.5: The classification accuracy of the language streams for each class of the RVL-CDIP dataset.

| Model | Adv. | Budg. | Email | File | Form | Handw. | Inv. | Letter | Memo | News | Pres. | Quest. | Res. | Public. | Report | Spec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GloVe | 0.53 | 0.68 | 0.85 | 0.90 | 0.62 | 0.53 | 0.81 | 0.57 | 0.62 | 0.78 | 0.56 | 0.72 | 0.94 | 0.77 | 0.62 | 0.85 |
| FastText | 0.57 | 0.72 | 0.89 | 0.94 | 0.68 | 0.64 | 0.88 | 0.69 | 0.70 | 0.78 | 0.62 | 0.81 | 0.95 | 0.85 | 0.73 | 0.88 |
| $\text{Bert}_{Base}$ | 0.68 | 0.83 | 0.95 | 0.85 | 0.80 | 0.69 | 0.88 | 0.84 | 0.90 | 0.84 | 0.82 | 0.87 | 0.97 | 0.89 | 0.80 | 0.92 |

Table 2.6: The classification accuracy of the vision modalities for each class in RVL-CDIP dataset.

| Model | Adv. | Budg. | Email | File | Form | Handw. | Inv. | Letter | Memo | News | Pres. | Quest. | Res. | Public. | Report | Spec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Inception-ResNetv2 | 0.89 | 0.78 | 0.97 | 0.96 | 0.72 | 0.93 | 0.88 | 0.82 | 0.93 | 0.83 | 0.72 | 0.75 | 0.96 | 0.87 | 0.86 | 0.85 |
| $\text{NasNet}_{Mobile}$ | 0.91 | 0.79 | 0.97 | 0.95 | 0.75 | 0.95 | 0.70 | 0.79 | 0.83 | 0.90 | 0.81 | 0.68 | 0.94 | 0.80 | 0.63 | 0.85 |
| $\text{NasNet}_{L_{4032d}}$ | 0.92 | 0.90 | 0.98 | 0.94 | 0.84 | 0.94 | 0.91 | 0.89 | 0.94 | 0.91 | 0.85 | 0.89 | 0.96 | 0.93 | 0.82 | 0.93 |
| $\text{NasNet}_{L_{768d}}$ | 0.94 | 0.90 | 0.98 | 0.96 | 0.83 | 0.95 | 0.93 | 0.90 | 0.93 | 0.92 | 0.85 | 0.89 | 0.96 | 0.93 | 0.82 | 0.93 |

categories (Form, Report, Email) shown in Figure 2.1. Finally, our cross-modal feature learning approach with superposing fusion outperforms all current state-of-the-art baselines with a significant margin of 3.91% compared to the two-stream-based methods, and of 2.53% compared to the single-modal-based methods. Thus, the superposing fusion (*i.e.* S.F) approach raises the performance of all classes regarding their structural property differences.

Out of the two proposed methods that merge both textual and visual features, the superposing fusion (*i.e.* S.F) method jointly learns more relevant information from textual features and visual features, achieving the best performance with 99.71% classification accuracy.

### Evaluation on the RVL-CDIP Dataset

To evaluate the effectiveness of our proposed cross-modal approach for document image classification, we firstly investigate the performance of the single-modal modalities based on the textual content and the corresponding visual features. As seen in Table 2.5, the classification results of each class of the three word embedding procedures are very low concerning three main categories that are: Advertisement, File Folder, and Handwritten. For Glove, the classification results of the three classes are 53%, 90%, and 53% respectively. Whereas for FastText, it improved slightly the accuracy results for each class to 57%, 94%, and 64% respectively. More specifically, the GloVe method predicted 36.32%,

Table 2.7: The classification accuracies of the cross-modal network for each class of the RVL-CDIP dataset, with the proposed fusion modalities.

| Model | Adv. | Budg. | Email | File | Form | Handw. | Inv. | Letter | Memo | News | Pres. | Quest. | Res. | Public. | Report | Spec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Concat. | 0.97 | 0.96 | 0.98 | 0.98 | 0.93 | 0.97 | 0.97 | 0.95 | 0.97 | 0.96 | 0.94 | 0.97 | 0.99 | 0.97 | 0.94 | 0.98 |
| Avg. | 0.97 | 0.97 | 0.98 | 0.98 | 0.94 | 0.97 | 0.97 | 0.95 | 0.97 | 0.96 | 0.94 | 0.97 | 0.99 | 0.97 | 0.95 | 0.98 |

32.66% of Advertisement and Handwritten class documents as File Folder documents. Also, FastText managed to improve the performance and reduced the classification error by 4% where 31.13% of Advertisement, and 28.28% of Handwritten class documents are predicted as File Folder documents. Furthermore, the bidirectional Bert$_{Base}$ enhanced the performance to 68% for Advertisement, 85% for File Folder, and 69% for Handwritten categories. The Bert$_{Base}$ network boosted the performance of the three classes and cut the error-classification by half where 15.98% of Advertisement, and 15.84% of Handwritten categories are predicted as File Folder document images. The classification errors are mainly due to either OCR error recognition, or empty document images which result to empty text files. Advertisement documents contain mostly images with few invisible text sequences, where the corresponding text generated by OCR is much too noisy and non-recognized. File Folder class presents in most cases empty document images with no text in it to be processed by the OCR engine. Finally, OCR technique fail to recognize handwritten characters in document images as a result of the different handwriting manners. Still, all vision networks trained on the RVL-CDIP dataset achieve comparable performance with the state-of-the-art methods. Table 2.6 illustrates the performance of our best single-modal vision modality NasNet$_{Large\ (768d)}$. It shows an improvement in the classification results of all classes, especially for the classes Advertisement, File Folder, and Handwritten to 94.08%, 96.04%, 95.07% in comparison of language modality results. Nevertheless, the lightweight NasNet$_{Mobile}$ network fails to improve the performance for most of the classes compared to Bert$_{Base}$, our best language-based model. Whereas, the Inception-ResNet-v2 network slightly outperforms our language modalities with 85.04% accuracy in comparison to Bert$_{Base}$ model (84.96%), surpassing significantly both Glove and FastText word embeddings.

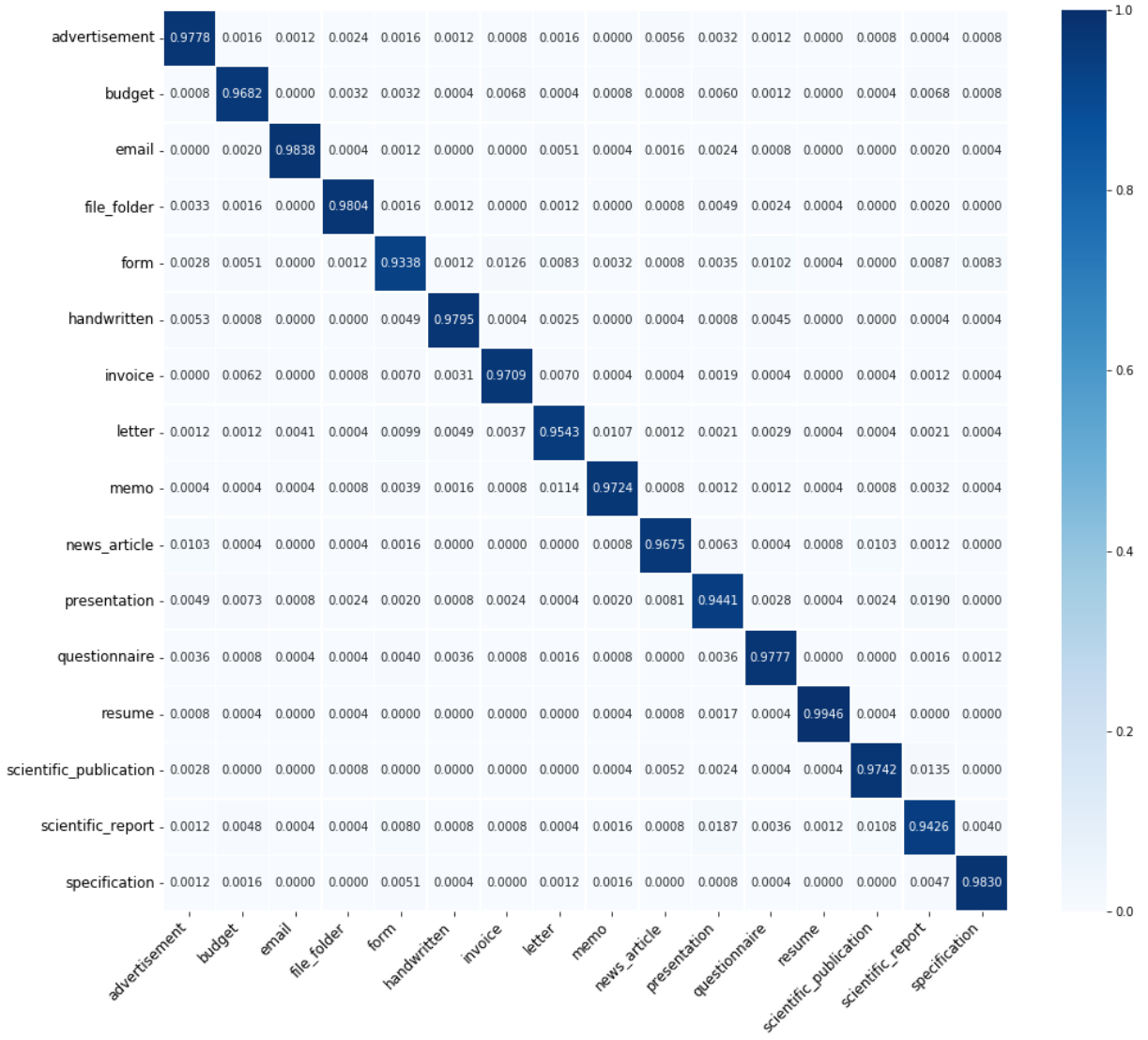



Figure 2.4: Confusion Matrix of the Equal Concatenation fusion scheme for the proposed cross-modal feature learning network.

Besides, the aim of this work is to leverage the ability of the cross-modal network to enhance the performance compared to the single-modal modalities. To do so, we proposed to merge textual and visual features with two different fusion modalities. For the superposing fusion method, it requires two feature vectors with the same size. Since the language output vector is of size 768, and the vision output vector is of size 4032, we added a fully connected layer on top of $\text{NasNet}_{Large\ (4032)}$. We re-trained it to study its effect on the classification results. Table 2.2 shows that indeed, adding a fully connected layer slightly increases the performance of the vision modality from 91.12% for $\text{NasNet}_{Large\ (4032)}$, to

91.45% for NasNet$_{Large\ (768)}$. This comparison illustrates that visual features are more important than textual features with both feature embeddings of size $4032d$ and $768d$.

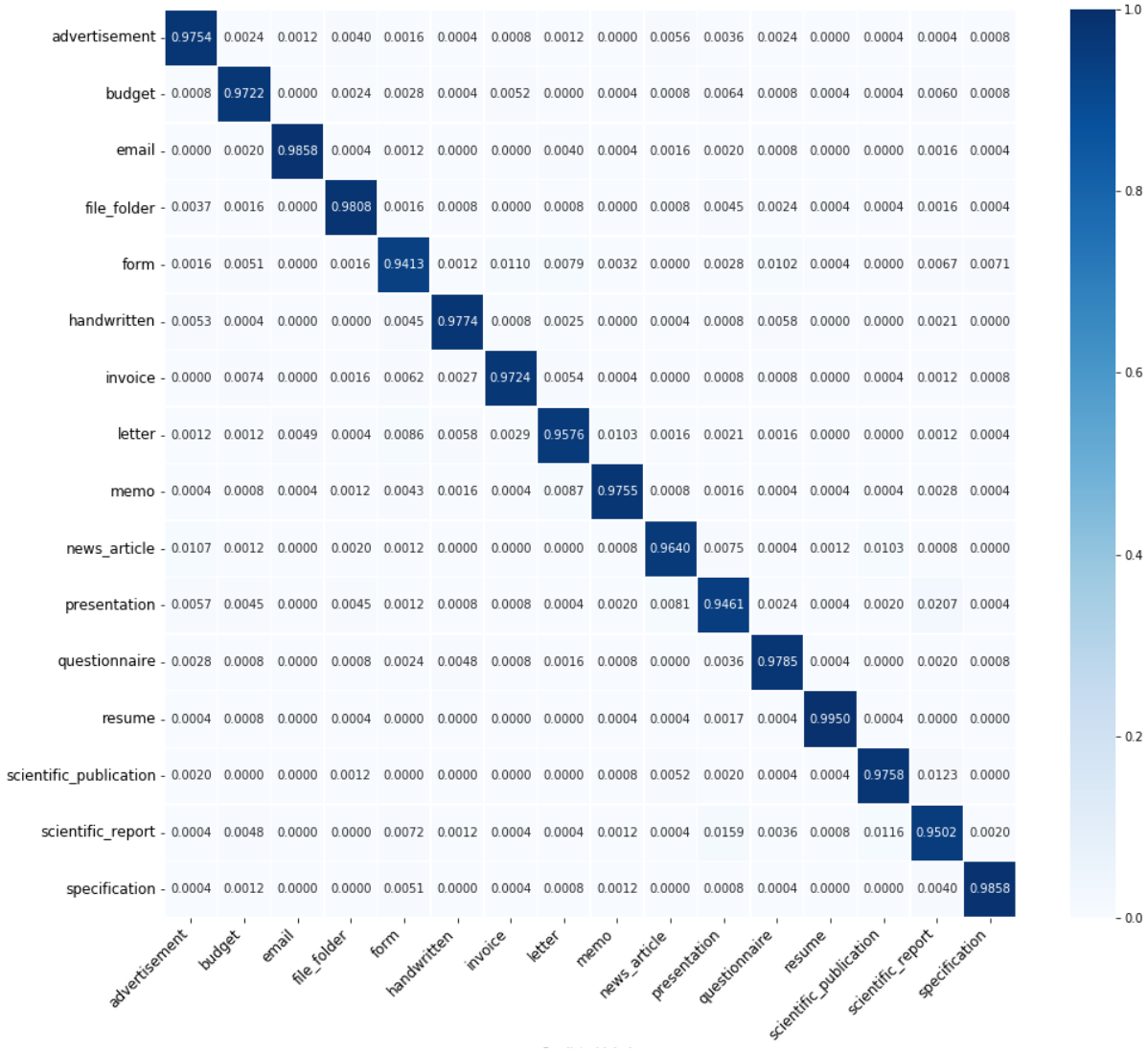



Figure 2.5: Confusion Matrix of our best cross-modal network with the superposing (*i.e.* A.E) fusion method.

Accordingly, Tables 2.2 and 2.7 show the accuracy of each class and the overall accuracy of the cross-modal network that merges the best single-modal modalities NasNet$_{Large}$, and Bert$_{Base}$. Jointly learning both modalities with an early fusion scheme achieves accurate results in comparison with the current state-of-the-art methods. The joint learning approach shows its capability to learn more relevant information from document images. Thus, it improves the accuracy of each class independently in comparison to single-modal

modalities. The cross-modal network manages to correct the error of the text classification generated by the language-based approaches for the three main classes: Advertisement, File Folder, and Handwritten.

Also, Figures 2.4 and 2.5 show the confusion matrices of our state-of-the-art cross-modal network with the equal concatenation and superposing fusion methods respectively. The network performs the best for the Resume category with a 99.46%, 99.50% classification accuracy for equal concatenation and superposing respectively. Whereas it performs the worst for the class Form with a 93.38%, 94.13% accuracy for the two fusion modalities.

To this end, we conclude that either Glove, FastText, or Bert are not able to outperform the vision-based approaches for this task. This proves that relying only on textual content is not sufficient. Hence, it needs the visual features to achieve accurate results. It is clear from all reported results that combining the visual structural properties of document images with the extracted text corpus improves the quality and accuracy of the final predictions for the document classification task.

### Additional Results

Table 2.8: The Recall and Precision metrics of the vision backbones of the most relevant classes in the RVL-CDIP dataset.

| Model | Metrics | Adve. | Email | Folder | Form | Hand. | Invoice | Pres. | Quest. | Resume | Sci. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\text{NasNet}_{Mobile}$ | Recall | 0.91 | 0.97 | 0.96 | 0.75 | 0.95 | 0.71 | 0.82 | 0.69 | 0.95 | 0.63 |
| | Precision | 0.83 | 0.90 | 0.89 | 0.69 | 0.83 | 0.95 | 0.68 | 0.87 | 0.82 | 0.77 |
| Inception-ResNet-v2 | Recall | 0.90 | 0.97 | 0.97 | 0.73 | 0.93 | 0.88 | 0.73 | 0.76 | 0.96 | 0.86 |
| | Precision | 0.90 | 0.99 | 0.86 | 0.78 | 0.93 | 0.83 | 0.84 | 0.90 | 0.91 | 0.56 |
| $\text{NasNet}_{Large_{4032d}}$ | Recall | 0.94 | 0.99 | 0.96 | 0.84 | 0.95 | 0.93 | 0.86 | 0.90 | 0.97 | 0.83 |
| | Precision | 0.92 | 0.98 | 0.95 | 0.86 | 0.95 | 0.93 | 0.84 | 0.87 | 0.98 | 0.83 |
| $\text{NasNet}_{Large_{768d}}$ | Recall | 0.93 | 0.99 | 0.95 | 0.84 | 0.92 | 0.92 | 0.85 | 0.89 | 0.97 | 0.82 |
| | Precision | 0.93 | 0.98 | 0.96 | 0.84 | 0.95 | 0.94 | 0.82 | 0.88 | 0.97 | 0.83 |

As illustrated in the Table 2.8, the lightweight $\text{NasNet}_{Mobile}$ framework fails to capture higher level features from Form, Invoice, Questionnaire, and Scientific report classes. The model seems to be less sensitive with a recall rate of 75%, 71%, 69%, and 63% for the four classes respectively. Also, we measured the precision of the $\text{NasNet}_{Mobile}$ network for

Table 2.9: The Recall and Precision metrics of the vision backbones of the most relevant classes of the RVL-CDIP dataset.

| Model | Metrics | Adve. | Email | Folder | Form | Hand. | Invoice | Pres. | Quest. | Resume | Sci. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Glove | Recall | 0.54 | 0.86 | 0.91 | 0.62 | 0.54 | 0.81 | 0.57 | 0.73 | 0.95 | 0.63 |
| | Precision | 0.61 | 0.88 | 0.41 | 0.81 | 0.57 | 0.80 | 0.64 | 0.93 | 0.97 | 0.63 |
| FastText | Recall | 0.57 | 0.90 | 0.94 | 0.69 | 0.64 | 0.88 | 0.63 | 0.82 | 0.95 | 0.74 |
| | Precision | 0.77 | 0.96 | 0.45 | 0.85 | 0.60 | 0.80 | 0.76 | 0.89 | 0.99 | 0.70 |
| $\text{Bert}_{Base}$ | Recall | 0.68 | 0.95 | 0.86 | 0.80 | 0.69 | 0.88 | 0.82 | 0.87 | 0.98 | 0.80 |
| | Precision | 0.78 | 0.97 | 0.60 | 0.81 | 0.83 | 0.90 | 0.81 | 0.89 | 0.99 | 0.82 |

each class. It is less precise with a precision rate of 68%, 69% for the classes Presentation and Form. Furthermore, the Inception-ResNet-v2 framework's recall rate for the classes Form, Presentation, and Questionnaire is low in comparison with other categories. The recall for each class is of 73%, 73%, and 76% respectively, while the precision is of 78% for the class Form, with a deterioration to 56% for the class Scientific report.

Lastly, for our best heavyweight model $\text{NasNet}_{Large}$, it shows an important ability to classify document images with a lower recall and precision of 83% for the Scientific report category. The higher recall is of 99% for the class Email, while the higher precision is of 98% for both Email and Resume classes. On the other hand, Table 2.9 illustrates the importance of $\text{Bert}_{Base}$ in capturing meaningful information, thus, improving the recall and precision rates for each category of the RVL-CDIP dataset compared to other word embedding models.

## 2.5 Discussion

In this chapter, we proposed a cross-modal methodology that learns simultaneously from the input token embeddings extracted from the text corpus, and the structural information from document images to perform document image classification. We showed that, merging the two modalities with different fusion schemes boosts the performance compared to single-modal networks. The dynamic $\text{Bert}_{Base}$ word embedding has proved its efficiency to learn relevant semantic information from the text corpus compared to static word embeddings, as well as the ability of heavyweight networks to learn higher level features compared

to lightweight architectures. The extensive experimental results achieved state-of-th-art performance on the two benchmark RVL-CDIP and Tobacco-3482 document datasets.

After analyzing of the obtained single-modal results regarding the performance of the different backbones on document classification, we have seen that the accuracy regarding some document categories remains very low. Even-though we achieved compelling results with the early feature fusion scheme, we intended to learn better information with the goal to improve the classification performance of the specific document categories where the accuracy remains very low. For example, the accuracy of handwritten documents is high for the vision modality and very low for the language modality. In this case, a framework which can transfer the knowledge from one modality to another is needed to help both vision and language modalities to learn better information, and thus, improve representation learning. Meanwhile, some type of documents such as File Folder category do not contain any visual spatial information, in which case a stronger emphasis on the textual information within the language cues is highly required. Hence, the general idea of Chapter 3 is to improve the representation learning and the performance for the single-modal modalities in an intermediate/middle-fusion methodology, instead of the early fusion schemes adopted in the current chapter.

# CHAPTER 3

# Multimodal Deep Mutual Learning

Competition has been shown to be useful up to a certain point and no further,
but cooperation, which is the thing we must strive for today,
begins where competition leaves off.
– *Franklin D. Roosevelt*

## 3.1 Motivation

In this chapter, our goal is to improve the robustness of the proposed model in Chapter 2 for the task of document image classification. Instead of leveraging the visual and textual features through an early feature fusion methodology as in Chapter 2, we aim to learn higher-level interactions between the middle blocks of the vision and language modalities in an intermediate/middle-fusion fashion before fusing them in the final stage in an early fusion manner. In contrast to Chapter 2 where the learning process of the vision and language modalities is independent one from another, we intend in this chapter, to improve representation learning of single-modal modalities by transferring the knowledge from one modality to another during the training stage. Therefore, the cross-modal representations

will be improved accordingly. Thus, we aim to answer the following research question of **how to effectively coordinate, learn the connections, and model the interactions between vision and language modalities in a fully-supervised learning paradigm.** Meanwhile, we demonstrate in this chapter the generalization ability of deep networks to classify unseen document data. In a general overview, people learning new concepts can often generalize successfully from just a single example, yet machine learning algorithms typically require hundreds or thousands of examples to perform with similar accuracy [93]. People can also use learned concepts in richer ways than conventional algorithms for action, imagination, and explanation. This opens to the research question: **On a challenging document image classification task, are these multimodal interactions and alignment between visual and textual information sufficient to generalize the learned knowledge to the unseen document data as human-performance ?**

Multimodal methods for document classification rely mainly on vision and language modalities. They contain two or an ensemble of deep networks which are trained on large-scale datasets to extract discriminate features from the input data. With such approaches, the learning process of the vision modality and the language modality is still independent one from another. The output features of both modalities are subsequently combined together to perform an ensemble trainable document classification network [11, 48, 184, 185]. Yet, these independent learning approaches might be enhanced if the visual and the textual features share some mutual information between them.

In general, knowledge transfer-based approached have been extensively studied in the literature in the CV and NLP fields [14, 65, 131, 147]. These approaches encourage collaborative learning between modalities, allowing vision and language modalities to simultaneously learn their discriminant features in a mutual learning manner [203]. They aim to align the current modality to the other modality by minimizing the difference in class probabilities produced by each modality [65]. However, rather than the conventional distillation-based teacher-student approach with one-way knowledge transfer from a pre-trained teacher to a student [65], the mutual learning strategy starts with a pool of untrained students in a student-to-student peer-teaching model to learn to solve the

tasks collaboratively [203]. It turns out that conventional mutual learning achieves better results than independent learning in either a supervised or a conventional distillation learning approach from a larger pre-trained teacher. Nonetheless, conventional mutual learning is a bi-directional knowledge transfer-based method, in which the current student modality can learn from a better example from the other modality, meanwhile the good student learns from the worst modality. That is to say, if the other student is worse than the current student, then the negative knowledge will be introduced and might weaken the ongoing training. This violates the motivation of the conventional mutual learning setting.

Therefore, we encourage mutual learning by transferring the positive knowledge between vision and language modalities during the training stage. This constraint is realized by adding a truncated-Kullback–Leibler divergence loss (Tr-KLD$_{Reg}$) as a regularization term to the conventional supervised setting., which will be elaborated on the next sections. To the best of our knowledge, this is the first time to leverage a mutual learning approach along with a self-attention-based fusion module to perform document classification.

As we mentioned before, deep Learning has provided compelling results in various document understanding problems such as document classification, form understanding, receipt understanding, etc. Existing works covered several techniques including document binarization [2, 133], layout analysis [132, 155], and structural similarity constraints [30] for many document analysis tasks. However, to ensure a good generalization, many deep neural networks with large amount of parameters have been used for document classification in order to extract the most relevant visual features [98].

Unlike the general images from the ImageNet dataset [150], document images have a distinct visual style. Therefore, numerous studies on document processing tasks have used transfer learning. It has been shown to be effective in boosting the classification performance of document images [3, 36, 61], whereas randomly initialized networks are under-performing [76]. Additionally, from the perspective of a natural language processing classifier, document images can be categorized into various classes based on their textual content processed by an Optical Character Recognition (OCR) system [140, 170]. Yang *et al.* [189] presented a neural network to extract semantic enriched information from textual

content based on a word embedding mechanism. Also, Appiani *et al.* [10] described a system that exploits a structural analysis approach to characterize and automatically index heterogeneous documents with variable layout, by determining the class of the document image based on reliable automatic information extraction methods.

Nevertheless, the challenge of document images remains in their wide range of visual variability, where documents from the same category might have different spatial properties. Due to their various visual styles, relying on deep convolutional networks to extract visual properties to perform document image classification might fail to distinguish between highly correlated classes. The inter-class discrimination of document images might be smaller than the intra-class variability, where two or multiple document images of different categories can be visually, and in terms of their textual content, closer than two or multiple documents from the same category. This level of intra-class variability can be mitigated by introducing the latent semantic information from the text corpus within the document image. Once the visual features of the vision modality and the textual features of the language modality are extracted, they are leveraged into a multimodal network to combine both feature vectors into one feature vector based on a feature fusion methodology [12, 38, 126].

Thus, we introduce a mutual learning approach based on a truncated-Kullback–Leibler divergence regularization term (Tr-KLD$_{Reg}$). This approach enables the current modality to learn only the positive knowledge from the other modality and prevents the negative knowledge from being introduced in the ongoing learning of the current modality. The proposed mutual learning approach with regularization improves the quality of the final predictions of the single-modal and cross-modal modalities, and helps to overcome the drawback of the conventional mutual learning trained with the standard Kullback–Leibler divergence (KLD).

Furthermore, as one of the goals of this chapter is to combine visual and language features through a better multimodal feature fusion methodology, we introduce a self-attention-based feature fusion module that serves as a middle block in our ensemble trainable network. Moreover, we aim to simultaneously extract more powerful and representative features from different middle blocks of the vision and language modalities through the

self-attention-based feature fusion module. This approach enables us to focus more on the salient parts of the feature maps of each modality, and aims to capture relevant semantic information between the pairs of image regions and text words. Such self-attention-based modules have recently became an elemental component in many multimodal tasks such as visual question answering, image captioning, image-text matching, etc [79, 100, 125, 186]. Furthermore, we adopt an early average ensemble fusion scheme in the final model to ensure a more stable and better-performing solution for the task of document image classification.

This work is built on the results and analysis of Chapter 2. In the following parts, we denote mutual learning trained with the standard (KLD) as $\text{ML}_{KLD}$, mutual learning trained with regularization as $\text{ML}_{Tr-KLD_{Reg}}$, and ensemble self-attention-based mutual learning with regularization as $\text{EAML}_{Tr-KLD_{Reg}}$.

Following are the main contributions of this chapter:

- We introduce a mutual learning strategy with a regularization term to overcome the drawback of the conventional mutual learning. This approach allows the current modality in process to learn the positive knowledge from the other modality, instead of the negative knowledge which weakens the learning capacity of the current modality in process.

- We present a self-attention-based feature fusion module for a better multimodal feature extraction to perform fine-grained document image classification. Our proposed self-attention-module enhances the overall accuracy of the ensemble network and achieves state-of-the-art classification performance compared to single-modal and multimodal learning methods.

- We perform a comprehensive ablation study on the benchmark RVL-CDIP and Tobacco-3482 datasets to analyze the effectiveness of our proposed ensemble trainable network with/without the mutual learning approach, and with/without the self-attention-based feature fusion module.

- We evaluate the performance and the generalization ability of the proposed ensemble network on unseen document data through inter-dataset and intra-dataset evaluation

on the benchmark RVL-CDIP and Tobacco-3482 datasets for the single-modal and multimodal fusion modalities.

## 3.2 Approach

In this section, we present in detail the proposed multimodal mutual learning and self-attention-based feature fusion approaches.

The proposed ensemble deep network (see Figure 3.1) is based on a multimodal architecture, which consists of vision, language, and vision-language fusion modalities. The vision and language modalities are dedicated to extracting visual features and textual embeddings respectively. The fusion branch is used to combine the extracted visual and language features into multimodal features. After the training of the ensemble network, the classification of document images is conducted by either the vision modality or the language modality. Moreover, the visual features and the the text embeddings learned are fused to conduct document image classification in a multimodal fashion.

### 3.2.1 Vision Modality

The vision modality extracts the visual features using the Inception-ResNet-V2 [164] as a backbone network, which is a convolutional neural network that achieved state-of-the-art results on the ILSVRC image classification benchmark. The model has $54.36M$ parameters.

### 3.2.2 Language Modality

Further, we process all document images with an off-the shelf optical character recognition (OCR) system, *i.e.* Tesseract OCR[1] to extract the text from document images. Since document images from RVLCDIP and Tobacco-3482 datasets are well-oriented and relatively clean, it is quite straightforward to run the Tesseract OCR engine on such documents. We utilized this OCR engine to conduct a fully automatic page segmentation without orientation or script detection. We analyzed the output of the OCR and found a lot of errors in

[1] https://github.com/tesseract-ocr/tesseract

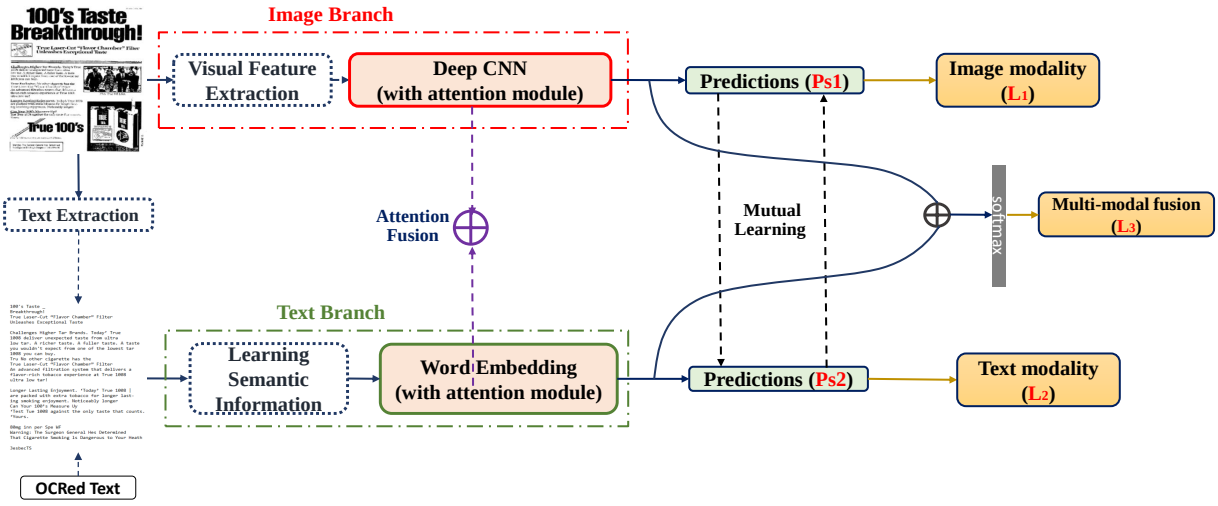



Figure 3.1: The proposed Ensemble Self-Attention-based Mutual Learning Network (EAML [18]).

the recognition especially for the classes Handwritten, and Notes, due to its incapability of recognizing handwriting. Besides, the Tesseract OCR engine is not always good at analyzing the natural reading order of documents. For example, it may fail to recognize that a document contains two columns, and may try to join text across columns, which is the case of some samples from the classes ADVE, and Scientific as shown in the qualitative results of the OCR engine in Figure 3.2. In addition, it may produce poor quality OCR results, as a result of poor quality scans of documents, or the distinct forms of document images as shown in Figure 3.2. They may contain handwritten text, tables, figures, and multi-column layouts. The embedded features extracted from the generated text corpus are computed using a Bert-based model [41]. It is a contextualized bi-directional word embedding mechanism, that uses a joint word representation conditioned on both left and right context in all layers using self-attention-based approaches.

### 3.2.3 Cross-Modal Modality

After the training of the vision modality/branch and the language modality/branch by the proposed mutual learning approach with regularization (*i.e.* $\text{ML}_{Tr-KLD_{Reg}}$), we attempt to fuse these two modalities/branches to simultaneously learn the visual and textual features extracted from the two vision and language branches. Moreover, we adopt an early fusion

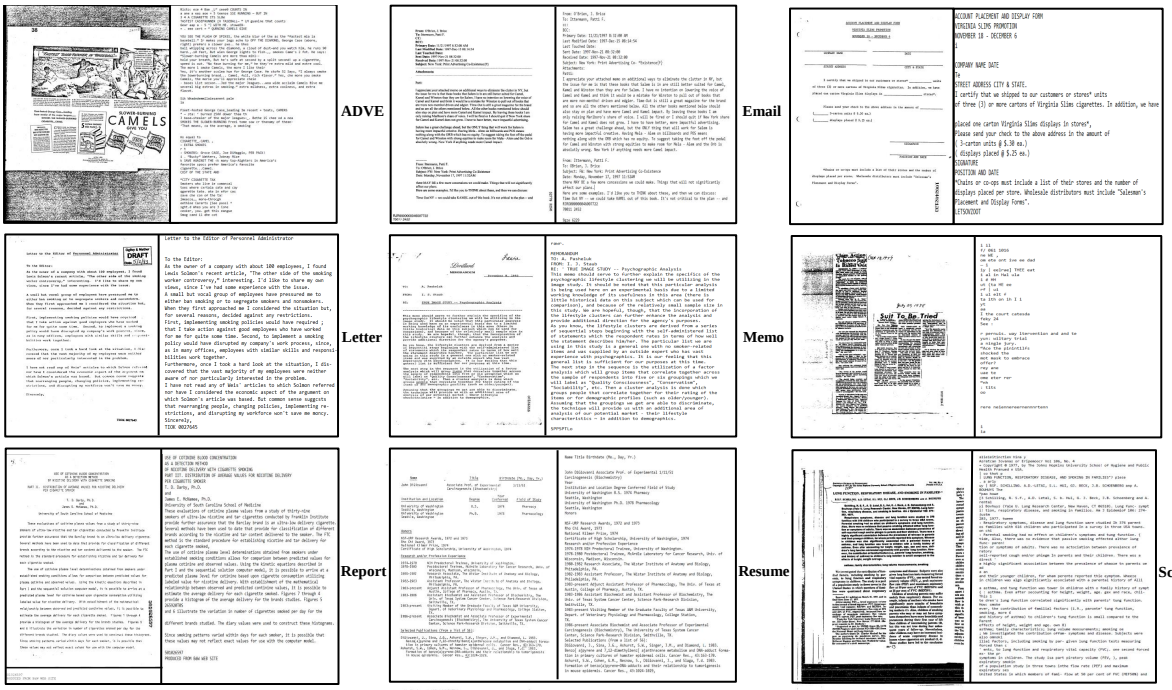
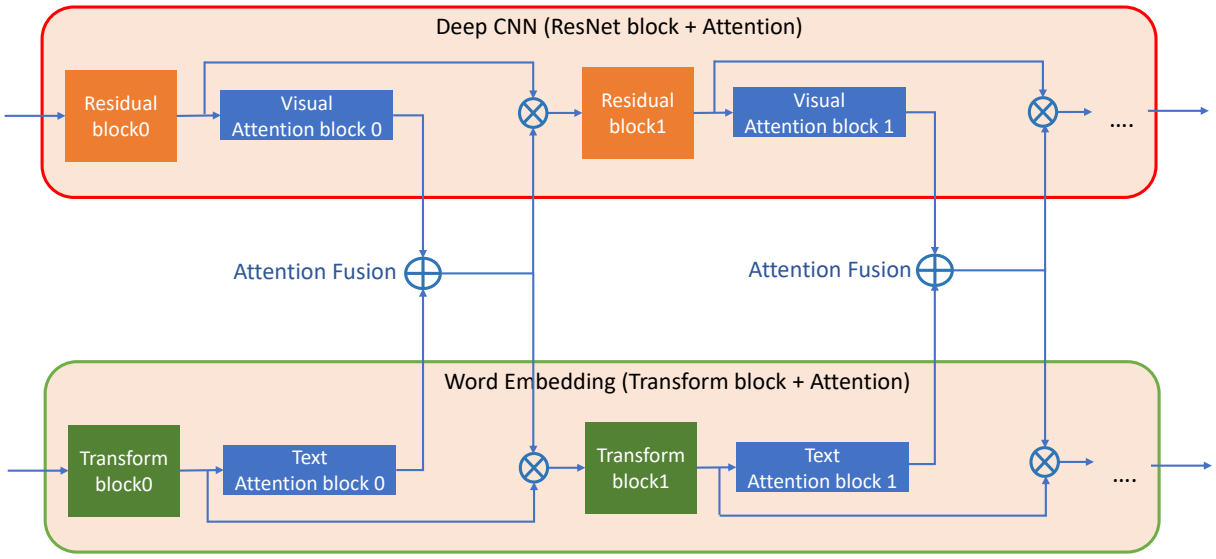



Figure 3.2: Sample document images and their corresponding OCR results of 9 classes of the Tobacco-3482 dataset that overlap with the RVLCDIP dataset.

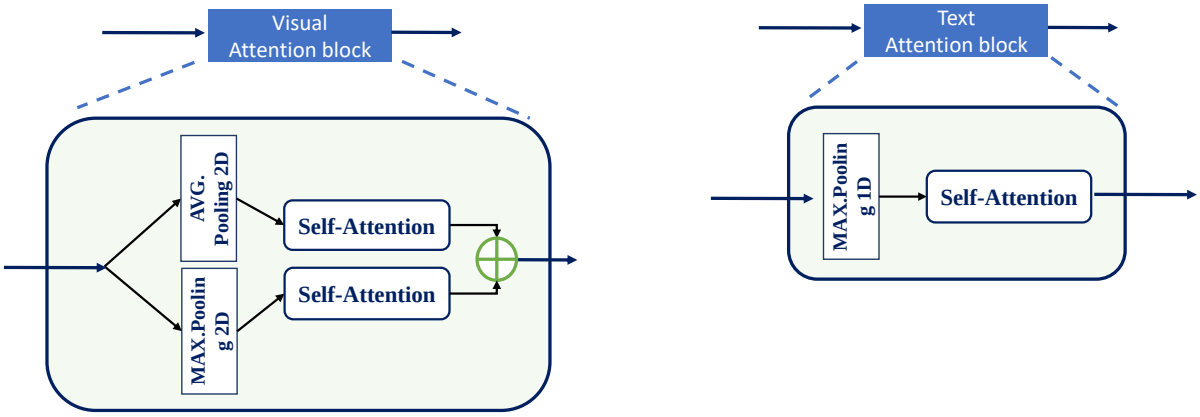
methodology, (*i.e.* average ensembling) as in Chapter 2, which enables us to enhance the global performance of multimodal networks.

### 3.2.4 Self-Attention-based Fusion Module

The proposed self-attention-based fusion module has been inspired by the attention modules in the squeeze and excitation network [67], which is based on re-weighting the channel-wise responses in a certain layer of a CNN by using soft self-attention in order to model the inter-dependencies between the channels of the convolutional features. As shown in Figure 3.3a, the attention fusion module is used as a middle fusion block in our ensemble trainable network. The intermediate features extracted from the middle blocks of the image branch (*e.g.* the output of Residual block0) and the text branch (*e.g.* the output of Transform block0) are passed to the corresponding attention block as the inputs of the attention block. The channel-wise information is then extracted from the input image or text intermediate features by performing down-sampling with the global average pooling and global max pooling layers in the attention blocks (see Figure 3.3a). The generated




(a) The architecture for the attention fusion module between vision modality and language modality [18].




(b) Visual/Textual attention blocks.

Figure 3.3: The proposed self-attention-based Fusion Module.

channel-wise features are then inputted to the self-attention block(s) to compute the attention maps. Specially, the self-attention maps obtained from the different self attention blocks are concatenated as the final self-attention map in the visual attention block.

Finally, the obtained self-attention maps from the visual attention block and text attention block are concatenated to generate the fusion attention map of the different modalities. The obtained fusion attention map is multiplied by the visual and textual intermediate features respectively (*i.e.* the input of the visual and text attention block) as the input to

the following Residual/Transform block in the image/text branch (see Figure 3.3b).

## 3.3 Proposed Method

In this section, we present in detail the proposed multimodal mutual learning and self-attention-based feature fusion approaches.

### 3.3.1 Multimodal Mutual Learning

As seen in the Figure 3.1, the proposed multimodal mutual learning network consists of three different modalities: vision modality (image branch), language modality (text branch) and the multimodal modality (fusion of the two vision an language modalities).

Consider a training dataset with a set of samples and labels $(x_n, y_n) \in (\mathcal{X}, \mathcal{Y})$, over a set of $K$ classes $\mathcal{Y} \in \{1, 2, .., K\}$. To learn the parametric mapping function $f_s(x_n) : \mathcal{X} \mapsto \mathcal{Y}$, we train our ensemble network with the parameter $f_s(x_n, \Theta)$, where $\Theta$ are the parameters obtained by minimizing a training objective function $\mathcal{L}_{train}$ denoted as:

$$\Theta = \arg\min_{\theta} \mathcal{L}_{train}(y, f_s(x, \theta)) \tag{3.1}$$

The total training loss of the ensemble network $\mathcal{L}_{train}$ is the sum of the weighted losses of the different modalities, *i.e.* the vision modality loss $\mathcal{L}_1$, the language modality loss $\mathcal{L}_2$ and the multimodal fusion (image/text) loss $\mathcal{L}_3$. Specifically, $\mathcal{L}_1$ and $\mathcal{L}_2$ are obtained by the mutual learning, which can be also called as the mutual learning loss. Thus, the total loss $\mathcal{L}_{train}$ for a pair $(x_n, y_n)$ is defined as follows:

$$\mathcal{L}_{train}(\mathbf{X_n}; \Theta) = \sum_{i=1}^{M} w_i \mathcal{L}_i(\mathbf{X}_n^{(i)}; \Theta_i) = w_1\mathcal{L}_1 + w_2\mathcal{L}_2 + w_3\mathcal{L}_3 \tag{3.2}$$

where $M = 3$ is the number of modalities to be performed. $X_i$ and $\Theta_i$ are the corresponding features and the parameters learned from each modality, $\Theta = \{\Theta_i\}_{i=1}^{M}$ are the overall parameters of the networks to be optimized by $\mathcal{L}_{train}$. $w_i \in [0, 1]$ s.t. $\sum w_i = 1$ denote hyper-parameters which balance the independent loss terms. Thus $\mathbf{X}_i \in \mathbb{R}^{d_i}$, where $d_i$ is

the dimension of the features $X_i$, and $\mathcal{L}_i, w_i \in \mathbb{R}^1$.

### Mutual Learning Loss

The conventional mutual learning task loss consists of two losses: a supervised learning loss (*e.g.* cross-entropy loss) and a mimicry loss (*e.g.* Kullback-Leibler divergence (KLD). The conventional mutual learning setting aims to help the training of the current modality by transferring the knowledge between one or an ensemble of modalities in a mutual learning manner as in [203]. However, the knowledge learned from the other modality through the conventional (KLD) includes both the negative part and the positive part that is transferred to the current modality. Yet, instead of using the standard (KLD) in the original mutual learning [203], we propose a so-called truncated-KLD loss (Tr-KLD$_{Reg}$) as a new regularization term in the training loss of the current modality, which enables us to filter the negative knowledge learned from the other modality, and only keep the knowledge being positive to the current modality. In this work, the cross-entropy loss $\mathcal{L}_s$ of the current modality in the process can be written as:

$$\mathcal{L}_s(\mathbf{X};\Theta) = \sum_{k=1}^{K} -y_k \log(\mathcal{P}_s(\hat{y}_k|\mathbf{X},\theta_k)) \tag{3.3}$$

where the probability $\mathcal{P}_s$ is the softmax operation given by:

$$\mathcal{P}_s(\mathbf{X};\theta_k) = \frac{e^{f^{\theta_k}(\mathbf{X})}}{\sum_{k'}^{K} e^{f^{\theta_{k'}}(\mathbf{X})}} \tag{3.4}$$

where $K$ is the number of classes in the dataset, $y_k$ is the one-shot label of the feature $\mathbf{X}$ of the input sample, $P_s$ is the class probability estimated by the softmax function. The truncated-Kullback-Leibler divergence regularization (Tr-KLD$_{Reg}$) loss of the current modality in process $\mathcal{D}_{KL_{Reg}}$ is given by:

$$\mathcal{D}_{KL_{Reg}}(\mathcal{P}_{s_2} \parallel \mathcal{P}_{s_1}) = \sum_{k=1}^{K} \mathcal{P}_{s_2} \max\left\{0, \log\left(\frac{\mathcal{P}_{s_2}}{\mathcal{P}_{s_1}}\right)\right\} \tag{3.5}$$

where $P_{s_1}$ is the class probability estimated by the current modality, while $P_{s_2}$ refers to the class probability estimated by the other modality. In this way, the mutual learning approach transfers the positive knowledge learned from the current modality to the other modality, by adapting the conventional mutual learning with the constraints of the mimicry loss $\mathcal{D}_{KL_{Reg}}$ (*i.e.* Tr-KLD$_{Reg}$). In the following part, $P_{s_1}$ refers to the class probabilities of the vision modality, while $P_{s_2}$ refers to the class probabilities of the language modality.

(i) **Vision Modality Setting:** For the vision modality, the overall loss function $\mathcal{L}_1$ is given by:

$$\mathcal{L}_1(\mathbf{X}_1;\Theta_1) = \mathcal{L}_{s_1}(\mathbf{X}_1;\Theta_1) + \beta\mathcal{D}_{KL_{Reg}}(\mathcal{P}_{s_2} \parallel \mathcal{P}_{s_1}) \tag{3.6}$$

where $\beta = 0.5$ is a hyper-parameter denoting the regularization weight.

The motivation of the conventional mutual learning aims to augment the training capacity of the network, by introducing the mimicry loss to align the classification probability of the current modality to the other modality with better training. However, it is not always true that the other/language modality performs better than the current/vision modality. In that case, the ongoing training of the current/vision modality will be weakened by the sum of the mimicry loss with the supervised loss (*i.e.* the cross-entropy loss for the classification of the document image). For instance, the mutual learning with regularization $D_{KL_{Reg}}$ loss will encourage the current/vision modality to learn only the positive knowledge from the other/language modality, and thus, prevent the negative knowledge from being introduced in the ongoing training of the current/vision modality.

(ii) **Language Modality Setting:** For the language modality, the overall loss function $\mathcal{L}_2$ can be written as:

$$\mathcal{L}_2(\mathbf{X}_2;\Theta_2) = \mathcal{L}_{s_2}(\mathbf{X}_2;\Theta_2) + \beta\mathcal{D}_{KL_{Reg}}(\mathcal{P}_{s_1} \parallel \mathcal{P}_{s_2}) \tag{3.7}$$

Similar to the vision modality setting, the mutual learning with regularization $D_{KL_{Reg}}$ loss will prevent to transfer the negative knowledge that might be introduced from the

other/vision modality, and thus, will encourage the transfer of only the positive knowledge to the current/language modality throughout the training process.

#### Multimodal Learning Loss

Instead of classifying document images using the independent vision or language modalities mentioned before, we can also conduct document image classification in a multimodal manner by combining the visual features and textual embeddings extracted from the two modalities trained with the mutual learning approach with regularization (*i.e.* $\text{ML}_{Tr-KLD_{Reg}}$). We directly superpose the visual features of the trained vision modality and text embeddings of the trained language modality to generate the ensemble cross-modal features as shown in Equation 3.9. Note that the dimension of the features extracted from the vision modality and the language modality are equal in this work and are denoted as $d$. A softmax layer at the end of the network is used to learn the classification of document images based on the ensemble cross-modal features $\mathbf{X}_3$. The parameter $\Theta_3$ of the softmax layer is optimized by the cross-entropy loss function $\mathcal{L}_3(\mathbf{X}_3; \Theta_3)$ which is given by:

$$\mathcal{L}_3(\mathbf{X}_3; \Theta_3) = -\sum_{k=1}^{K} y_k log P(\hat{y}_k | \mathbf{X}_3, \Theta_3) \tag{3.8}$$

with $\mathbf{X}_3$ given by:

$$\mathbf{X}_3 = [X_1 + X_2], \quad \mathbf{X}_3 \in \mathbb{R}^d \tag{3.9}$$

### 3.3.2 Self-Attention-based Fusion Module

The aim of the self-attention-based fusion module (see Figure 3.3) is to enhance the representation of the concatenated visual and textual feature maps to capture their salient features while eliminating to some extent the irrelevant or noisy ones. The adopted self-attention-based fusion module has been inspired by the attention module in [67, 172], which is based on the channel-wise re-calibration of feature maps to model the dependency of channels. The intermediate feature maps of each individual modality can be

interpreted as a set of local descriptors that include global information in the decision process of the network. This is achieved by using global max pooling and global average pooling layers to generate channel-wise information. The advantage of these pooling operations is to enforce correspondences between feature maps and categories.

Consider a set of input features $\mathbf{X} = [\mathrm{x}_1, ..., \mathrm{x}_m] \in \mathbb{R}^{m.d_x}$ and output features $\mathcal{F} = [\mathrm{f}_1, ..., \mathrm{f}_m] \in \mathbb{R}^{m.d_f}$, where m is the number of samples, $d_x$ and $d_f$ are the dimensions of input and output features respectively. For the vision modality, the input features $\mathbf{X}$ are passed to global average pooling and global max pooling layers. The spatial information for each layer is computed as:

$$\mathcal{F}^{'}_{I_{Avg}} = GlobalAvgPool2D(\mathbf{X}_{I_{Avg}}) \tag{3.10}$$

$$\mathcal{F}^{'}_{I_{Max}} = GlobalMaxPool2D(\mathbf{X}_{I_{Max}}) \tag{3.11}$$

where $\mathcal{F}^{'}_{I_{Avg}}$, and $F^{'}_{I_{Max}}$ correspond to the intermediate feature maps of the intermediate input features $X_{I_{Avg}}$, and $X_{I_{Max}}$ of the vision modality.

For the language modality, the input features are fed to a global max pooling layer:

$$\mathcal{F}^{'}_{T_{Max}} = GlobalMaxPool1D(\mathbf{X}_{T_{Max}}) \tag{3.12}$$

where $F^{'}_{T_{Max}}$ corresponds to the intermediate feature maps of the input features $X_{T_{Max}}$ of the language modality.

For our proposed self-attention-based fusion module, the intermediate feature maps of the vision and language modalities extracted by the pooling operations are fed to three independent fully-connected layers which correspond to the vectors query, keys, and values respectively as follows:

$$\mathrm{Q} = \mathrm{FC}_q(\mathcal{F}^{'}); \tag{3.13}$$

$$\mathrm{K} = \mathrm{FC}_k(\mathcal{F}^{'}); \tag{3.14}$$

$$\mathrm{V}_I = \mathrm{FC}_v(\mathcal{F}^{'}) \tag{3.15}$$

where $\mathrm{Q}, \mathrm{K}, \mathrm{V} \in \mathbb{R}^{m.d}$ are three vectors of the same shape fed to the attention function, which consists of computing the compatibility of the query with the key vectors to retrieve the corresponding value.

Given a query q and all keys K, we calculate the dot products of q with all keys K, divide each by a scaling factor $\sqrt{d_f}$ and apply the softmax function to get the attention weights on the values. The output features of each self-attention module of vision and language modalities $\mathcal{F}$ are given as follows:

$$\mathrm{A} = \mathrm{softmax}(\frac{\mathrm{Q}.K^{\mathcal{T}}}{\sqrt{d_f}}) \tag{3.16}$$

$$\mathcal{F} = \mathrm{A}.\mathrm{V} \tag{3.17}$$

where A is the attention map containing the attention weights for all query-key pairs, and the output features of the self-attention blocks $\mathcal{F}$ are the weighted summation of the values V determined by the attention function A.

Learning an accurate attention map A is crucial for self-attention learning. The scaled dot-product attention in Equations 3.16 and 3.17 models the relationship between feature pairs. Once the spatial information is extracted and fed into the self-attention blocks to compute the attention maps, they are then concatenated and multiplied by the input features of the vision and language modalities for adaptive feature fusion, which is computed as follows:

$$\mathcal{M}(\mathcal{F}) = \sigma(\mathcal{F}).\mathcal{F} \tag{3.18}$$

where $\mathcal{M}$ is the feature map that is passed to the following intermediate vision and language blocks of the vision and language modalities. The term $\sigma(.)$ denotes the sigmoid function. This feature map generated by the proposed self-attention-based fusion module focuses on the important features of the channels and concentrates on where the salient features are located.

## 3.4 Experimental Setup

### 3.4.1 Preprocessing

As the vision modality requires document images of a fixed size as an input, we first downscale all images to $229 \times 229$ pixels. Intuitively, when training DCNNs, data augmentation has been shown to be effective for real-world image classification [87]. The training data is augmented by shifting it horizontally and vertically with a range of 0.1. Also, shear transform is applied with a range of 0.1. To improve regularization of our vision modality, cutout [42] is applied, which augments the training data by partially occluded versions of the existing sample images. On the other hand, document images from the RVL-CDIP dataset are well-oriented and relatively clean. Hence, we run the Tesseract OCR engine. We used the version $4.0.0 - beta.1$ of Tesseract based on a LSTM engine to aim for better accuracy. The resulting extracted text was not post-processed. Although document information might be lost in OCR, such as typeface, graphics, layout, stop words, mis-spellings, symbols and characters, it could benefit from some level of spell checking to improve the semantic learning. However, we chose to provide the true output of Tesseract OCR as is.

### 3.4.2 Implementation Details

The network used in our proposed approaches were conducted on a 4 NVIDIA RTX-2080 GPU, using stochastic gradient descent optimizer (SGD), with Nesterov momentum, mini-batch size of 16, and a learning rate of $1e-3$ decayed with a value of 0.5 every 10 epochs. the learning rate decay is defined as

$$lr = initial\_lr * drop^{\left(\frac{iter}{iter\_drop}\right)} \tag{3.19}$$

The mutual learning strategy with regularization (*i.e.* $\text{ML}_{Tr-KLD_{Reg}}$) is performed in each mini-batch throughout the training process. At each iteration, the predictions of each modality are computed and the parameters are updated according to the predictions of the other modality as in Equations 3.6, 3.7 and 3.8. The optimization process of parameters $\Theta_1$, $\Theta_2$, and $\Theta_3$ is performed iteratively until convergence. We considered

early stopping within 10 epochs to stop the training process once the model's performance stops improving on the hold out validation dataset.

## 3.5 Experiments and Ablation Study

### 3.5.1 Evaluation Protocol

To evaluate the performance and the generalization ability of our proposed ensemble network, we proceed with intra-dataset and inter-dataset evaluation on the benchmark RVL-CDIP and Tobacco-3482 datasets. For the intra-dataset evaluation, we train and test the model on the same dataset -in which the train set and test sets have the same data distribution- to evaluate the performance of the proposed approaches. Whereas, for the inter-dataset evaluation, we train and test the ensemble network on different datasets -having different data distribution- to evaluate the generalization ability of the trained model. We first train our ensemble network on the RVL-CDIP dataset, then we employ the intra-dataset evaluation on RVL-CDIP and the inter-dataset evaluation on Tobacco-3482. Secondly, we train our ensemble network on the Tobacco-3482 dataset, then we employ the intra-dataset evaluation on Tobacco-3482 and the inter-dataset evaluation on RVL-CDIP. Note that there is no overlap between training set and test set either in intra-dataset or inter-dataset evaluation.

We report the accuracy, recall, and precision metrics achieved on the test set for the following methods: Independent Learning based on the single-modal vision and language modalities; Mutual Learning trained with the standard Kullback-Leibler divergence (KLD); Mutual Learning trained with the truncated-Kullback-Leibler divergence regularization (Tr-KLD$_{Reg}$) loss; and Ensemble Self-Attention Mutual Learning trained with (Tr-KLD$_{Reg}$). We denote them respectively as IL, ML$_{KLD}$, ML$_{Tr-KLD_{Reg}}$, and EAML$_{Tr-KLD_{Reg}}$ (see Tables 3.2, 3.3). We also compute the average precision (AP) from prediction scores which summarizes a precision-recall curve as the weighted mean of precision achieved at each threshold, with the increase in recall from the previous threshold

used as the weight:

$$\mathrm{AP} = \sum_{n} (\mathrm{R}_n - \mathrm{R}_{n-1})\mathrm{P}_n \tag{3.20}$$

where $\mathrm{P}_n$ and $\mathrm{R}_n$ are the precision and recall at the $n^{th}$ threshold. The high area under the (AP) curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both precision and recall show that the model is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). In addition, we compare our work against other state-of-the-art methods on the RVL-CDIP and Tobacco-3482 datasets. Note that the baseline methods in Tables 3.1 and 3.4 are not necessarily based on vision and language modalities. For example, [185] leverages visual features to incorporate words' visual information into LayoutLM for document-level pre-training. Also, [184] leverages pre-training text, layout and image in a multimodal framework by using text-image alignment and text-image matching tasks in the pre-training stage, where the cross-modality interaction is better learned.

### 3.5.2 Intra-dataset Evaluation

#### Results on the RVL-CDIP Dataset

On the large-scale RVL-CDIP dataset, all of the adopted approaches in this work achieve comparable performance with the state-of-the-art models. We report the overall accuracy results in Table 3.1. compared to our previous results in Chapter 2 and in our work from [17] and other baseline methods. The proposed EAML$_{Tr-KLD_{Reg}}$ model achieves the best performance in terms of accuracy for the single-modal vision and language modalities, and for the multimodal fusion modality at an accuracy of 97.67%, 97.63%, and 97.70% respectively. The adopted self-attention-based fusion module has shown its effectiveness in capturing simultaneously the inter-modal interactions between image features and text embeddings, along with the mutual learning approach with regularization (*i.e.* ML$_{Tr-KLD_{Reg}}$). Therefore, it improves the global classification performance of the single-

Table 3.1: The overall classification accuracy of our best EAML$_{Tr-KLD_{Reg}}$ method against baseline methods on the RVL-CDIP dataset.

| Method | Model | Accuracy(%) |
|---|---|---|
| Vision | Nicolas *et al.* [12] | 89.1 |
| Language | | 74.6 |
| Multimodal | | 90.6 |
| Vision | Dauphinee *et al.* [38] | 90.24 |
| Language | | 82.23 |
| Multimodal | | 93.07 |
| Vision | Cross-Modal [17] | 91.45 |
| Language | | 84.96 |
| Multimodal | | 97.05 |
| Vision | EAML$_{Tr-KLD_{Reg}}$ (Ours) | **97.67** |
| Language | | **97.63** |
| Multimodal | | **97.70** |
| Baselines | Harley *et al.* [61] | 89.80 |
| | Csurka *et al.* [35] | 90.70 |
| | Tensmeyer *et al.* [166] | 90.94 |
| | Azfal *et al.* [4] | 90.97 |
| | Das *et al.* [36] | 91.11 |
| | Das *et al.* [36] | 92.21 |
| | Ferrando *et al.* [48] | 92.31 |
| | Xu *et al.* [185] | 94.42 |
| | Xu *et al.* [184] | 95.64 |

modal and cross-modal modalities and outperforms the state-of-the-art baselines.

### Evaluation of the Single-Modal Tasks on the RVL-CDIP dataset

(I.) **IL vs ML$_{KLD}$:** The reported results in Table 3.2 illustrate the impact of training the independent vision and language modalities in a mutual learning manner, on the learning process of both modalities. We observe that the ML$_{KLD}$ approach improves the classification performance of the vision modality from 85.04% to 88.87%, while it deteriorated the performance of the language modality from 84.96% to 80.89%. We explain this performance deterioration of the language modality by learning the negative knowledge from the vision modality. In fact, the knowledge transferred via the standard (KLD) loss harms the ongoing training of the current/language modality in process. Here, given image features from an image sample with its corresponding text embeddings, the negative learning

Table 3.2: The overall classification accuracy(Acc.), recall(R.), precision(Pr.) metrics of the proposed approaches on the RVL-CDIP dataset. IL, $\text{ML}_{KLD}$, $\text{ML}_{Tr-KLD_{Reg}}$, and $\text{EAML}_{Tr-KLD_{Reg}}$ denote Independent Learning, Mutual Learning with the standard KLD, Mutual Learning with the truncated-KLD, and Ensemble self-attention-based Mutual Learning with the truncated-KLD respectively.

| | Modality | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Vision Modality | | | Language Modality | | | Cross-Modal Fusion | | |
| | Acc.(%) | R. | Pr. | Acc.(%) | R. | Pr. | Acc.(%) | R. | Pr. |
| IL | 85.04 | 0.85 | 0.85 | 84.96 | 0.85 | 0.85 | 94.44 | 0.94 | 0.94 |
| $\text{ML}_{KLD}$ | 88.87 | 0.89 | 0.88 | 80.89 | 0.81 | 0.80 | 90.06 | 0.90 | 0.90 |
| $\text{ML}_{Tr-KLD_{Reg}}$ | 90.81 | 0.91 | 0.91 | 88.80 | 0.89 | 0.89 | 96.28 | 0.96 | 0.96 |
| $\text{EAML}_{Tr-KLD_{Reg}}$ | **97.67** | **0.98** | **0.98** | **97.63** | **0.98** | **0.98** | **97.70** | **0.98** | **0.98** |

comes from the low class probabilities predicted by the vision modality, while at the same time, the language modality has made the right predictions from the same sample. In this way, the mutual training is harmed for the language modality and its loss variation $\mathcal{L}_2(\mathbf{X}_2;\Theta_2)$ becomes slower. Thus, using the Mutual Learning $\text{ML}_{KLD}$ approach actually makes the language modality worse than the Independent Learning (IL) approach.

Nonetheless, for the vision modality, the classification accuracy has improved. This means that transferring the knowledge from the language modality to the vision modality by learning mutually from the text predictions is effective.

(II.) **IL vs $\text{ML}_{Tr-KLD_{Reg}}$:** The classification results in Table 3.2 show that, training the vision and language modalities in a mutual learning manner -trained with the regularization term (*i.e.* $\text{Tr-KLD}_{Reg}$)- provide an improvement compared to the IL and the $\text{ML}_{KLD}$ methods. It improves the classification accuracy of the vision modality from 85.04% for the IL method to 90.81% for the $\text{ML}_{Tr-KLD_{Reg}}$ method. Also, it enhances the predictions of the language modality from 84.96% to 88.80% respectively.

Accordingly, the network keeps learning only from its cross-entropy loss $\mathcal{L}_s(\mathbf{X};\Theta)$ when the knowledge to be transferred from the other modality will harm the ongoing training of the current modality.

(III.) **$\text{ML}_{Tr-KLD_{Reg}}$ vs $\text{EAML}_{Tr-KLD_{Reg}}$:** The proposed self-attention-based fusion module for visual and textual feature fusion focuses on the salient feature maps gen-

erated from the image and the text modalities and suppresses the unnecessary ones to efficiently leverage these two modalities. The introduction of this attention module to fuse the two modalities along with the mutual learning approach has shown its efficiency compared to the $\text{ML}_{Tr-KLD_{Reg}}$ method as shown in Table 3.2. We demonstrate that the $\text{EAML}_{Tr-KLD_{Reg}}$ method outperforms $\text{ML}_{Tr-KLD_{Reg}}$ method with a significant margin at an accuracy of 97.67%, 97.63% for the vision and language modalities respectively. The attention module enhances the classification performance of all classes for the single-modal modalities. therefore, leveraging both modalities to one another in a middle fusion manner along with the mutual learning strategy encourage collaborative learning.

**Evaluation of the Multimodal Tasks on the RVL-CDIP Dataset**

In the multimodal learning task, the learned visual and language features are combined to conduct document image classification. At first, from Table 3.2, we see that the multimodal fusion predictions outperform the independent predictions of the single-modal modalities for each method. Moreover, jointly learning both modalities in an ensemble network benefit from training vision modality and text modalities both independently (IL) and in a mutual learning manner ($\text{ML}_{Tr-KLD_{Reg}}$). The ensemble predictions learned across the $\text{EAML}_{Tr-KLD_{Reg}}$ method with an accuracy of 97.70%, outperform the predictions learned from training the ensemble network across either the $\text{ML}_{Tr-KLD_{Reg}}$, the $\text{ML}_{KLD}$, or the IL approaches at an accuracy of 96.28%, 90.06%, and 94.44% respectively. That is to say, the ability of the self-attention-based fusion module along with the mutual learning strategy -trained with the regularization term (*i.e.* $\text{Tr-KLD}_{Reg}$)- to improve ensemble models is beneficial for the task of document image classification, which outperforms the state-of-the-art results for the multimodal task as seen in Table 3.1. Accordingly, the proposed $\text{EAML}_{Tr-KLD_{Reg}}$ method manages to correct the classification errors produced by vision and language modalities during the learning process. Hence, it provides state-of-the-art classification results for the task of document image classification.

In this manner, we showed the effectiveness of leveraging visual and textual features learned in a mutual learning with regularization strategy through a self-attention-based feature fusion module. Our approach learns simultaneously relevant and accurate infor-

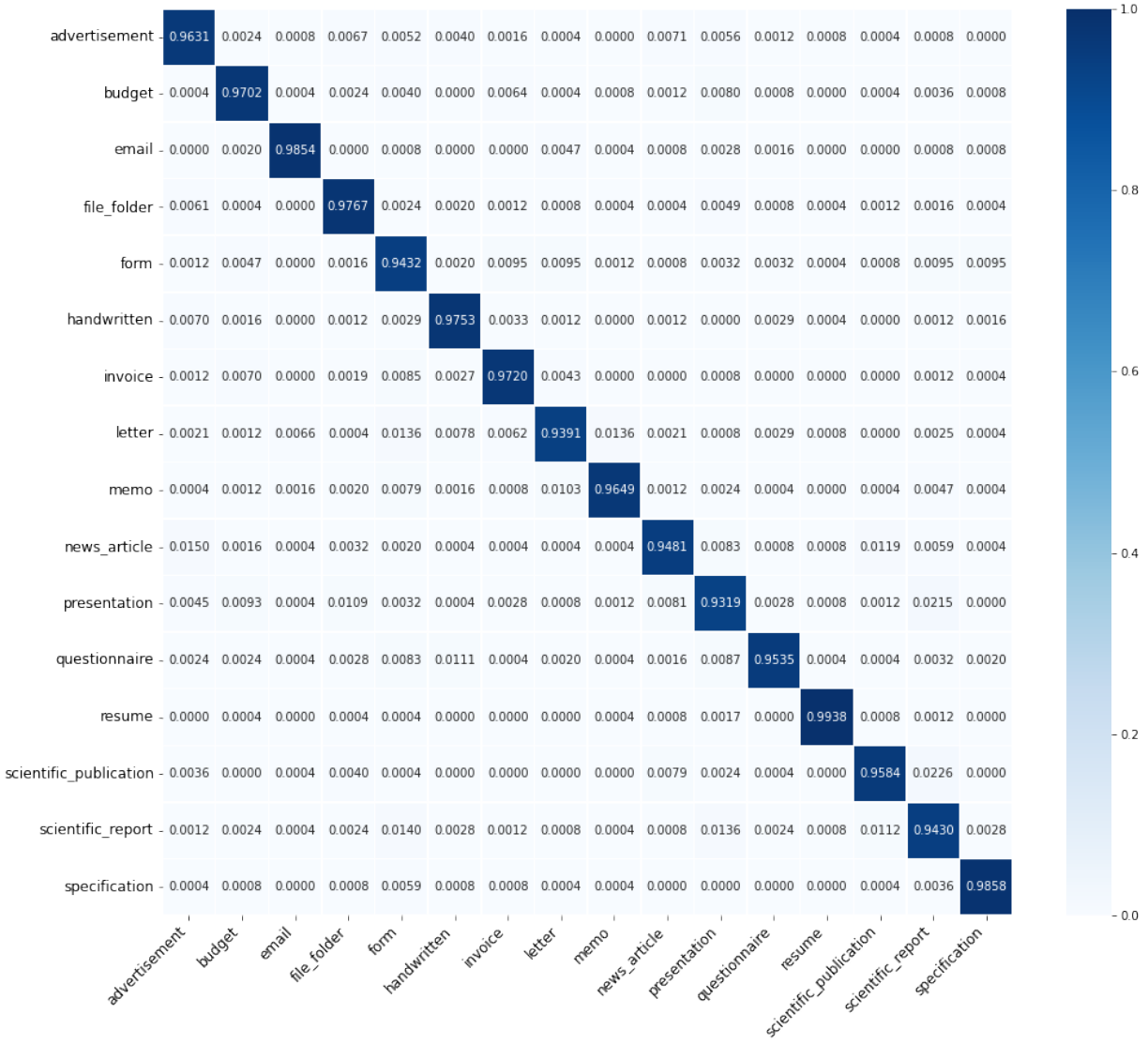



Figure 3.4: Multimodal Fusion Modality of the $\text{ML}_{Tr-KLD_{Reg}}$ method.

mation from the vision modality, and the language modality during the training stage. It enhances the ensemble model predictions by encouraging attention collaborative learning from one modality to another. Also, it boosts the overall classification performance. We report in Figures 3.4 and 3.5, the confusion matrices of the cross-modal modalities of the $\text{ML}_{Tr-KLD_{Reg}}$ and the $\text{EAML}_{Tr-KLD_{Reg}}$ methods respectively.

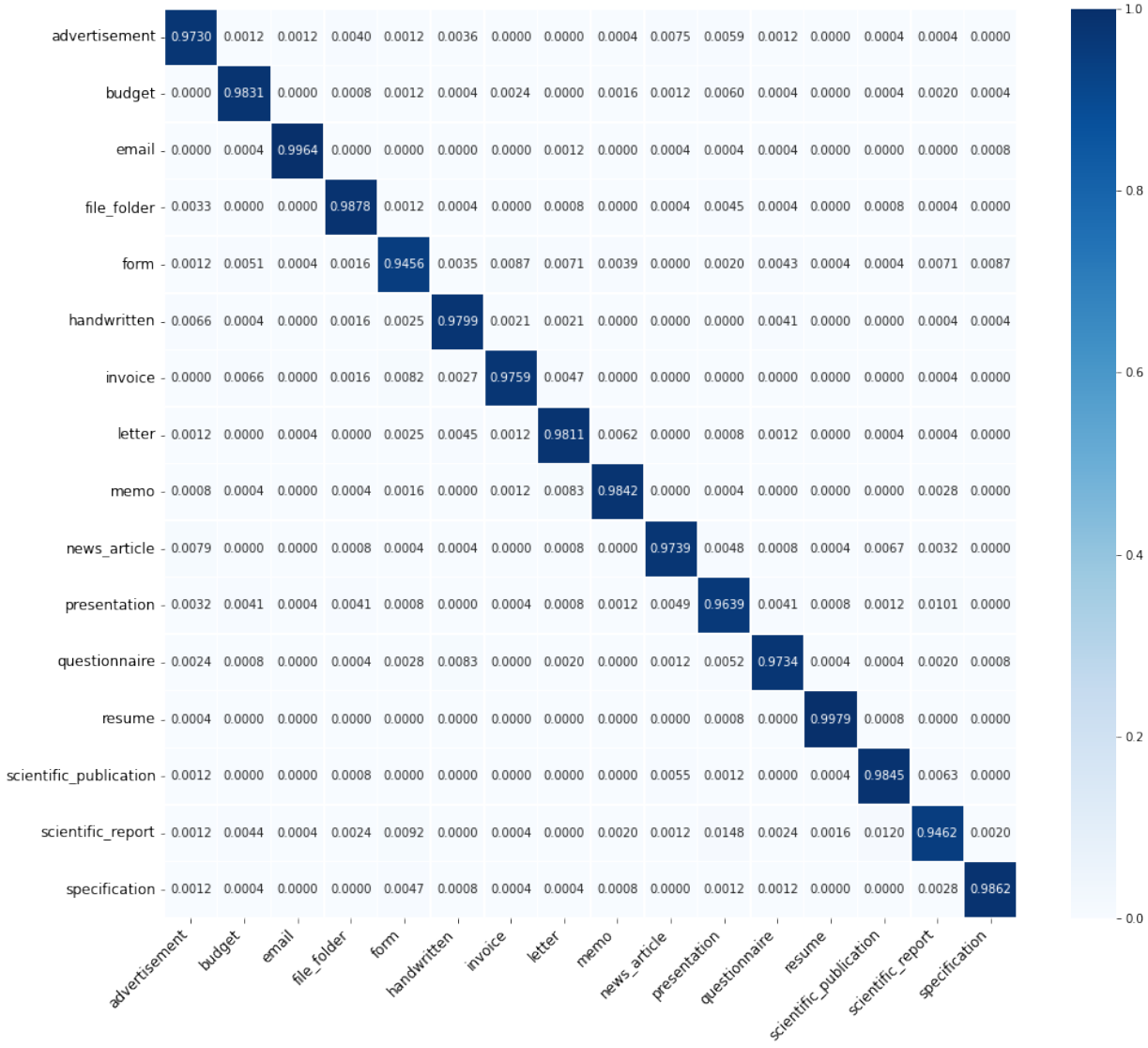



Figure 3.5: Multimodal Fusion Modality of the EAML$_{Tr-KLD_{Reg}}$ method.

### Results on the Tobacco-3482 Dataset

As reported in Table 3.3, which corresponds to the achieved performance on the Tobacco-3482 dataset, the EAML$_{Tr-KLD_{Reg}}$ method improves the classification performance significantly. The proposed EAML$_{Tr-KLD_{Reg}}$ method improves the overall performance of the single-modal and cross-modal modalities at an accuracy of 97.99%, 96.27%, and 98.57% for the vision modality, for the language modality, and for the multimodal fusion modality respectively compared to other methods. Thus, it achieves compelling performance results compared to the baseline methods on the Tobacco-3482 dataset (see Table 3.4).

Table 3.3: The overall classification accuracy(Acc.), recall(R.), precision(Pr.) metrics of the proposed approaches on the Tobacco-3482 dataset. IL, $\text{ML}_{KLD}$, $\text{ML}_{Tr-KLD_{Reg}}$, and $\text{EAML}_{Tr-KLD_{Reg}}$ denote Independent Learning, Mutual Learning with the standard KLD, Mutual Learning with the truncated-KLD, and Ensemble self-attention-based Mutual Learning with the truncated-KLD respectively.

| | Modality | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | vision modality | | | language modality | | | multimodal Fusion | | |
| | Acc.(%) | R. | Pr. | Acc.(%) | R. | Pr. | Acc.(%) | R. | Pr. |
| IL | 96.17 | 0.96 | 0.96 | 96.02 | 0.96 | 0.95 | 96.95 | 0.97 | 0.97 |
| $\text{ML}_{KLD}$ | 93.69 | 0.92 | 0.92 | 88.82 | 0.87 | 0.86 | 94.84 | 0.95 | 0.93 |
| $\text{ML}_{Tr-KLD_{Reg}}$ | 97.70 | 0.97 | 0.96 | 96.27 | 0.95 | 0.96 | 98.28 | 0.97 | 0.98 |
| $\text{EAML}_{Tr-KLD_{Reg}}$ | **97.99** | **0.97** | **0.98** | **96.27** | **0.95** | **0.96** | **98.57** | **0.98** | **0.98** |

Besides, the results illustrate that training the vision and language modalities in a mutual learning manner with the $\text{ML}_{KLD}$ method weakens the learning capacity of the language modality. Therefore, we show the effectiveness of the $\text{ML}_{Tr-KLD_{Reg}}$ approach that transfers only the positive knowledge from the current modality in process to the other modality.

### 3.5.3 Intra-Dataset Confusion Matrices

Figures 3.4 and 3.5 illustrate the confusion matrices of the $\text{EAML}_{Tr-KLD_{Reg}}$ and $\text{ML}_{Tr-KLD_{Reg}}$ methods. The figures show that the combined predictions from the vision and language modalities through a fusion methodology improve the classification accuracy of each class of the dataset independently, compared to the single-modal vision and language modalities. Furthermore, the $\text{EAML}_{Tr-KLD_{Reg}}$ method outperforms the $\text{ML}_{Tr-KLD_{Reg}}$ methods given the multimodal fusion classification results.

### 3.5.4 Inter-dataset Evaluation

This subsection describes an experimental investigation into the inter-dataset generalization of our fully-supervised deep network models, trained to distinguish between several categories of documents. The experiments conducted on inter-dataset evaluation question the implied link that learning cross-modal interactions and alignment between different

Table 3.4: The overall classification accuracy of the proposed approaches against baseline methods on the Tobacco-3482 dataset.

| Method | Model | Accuracy(%) |
|---|---|---|
| Image | Nicolas *et al.* [12] | 84.5 |
| Text | | 73.8 |
| multimodal | | 87.8 |
| Image | Asim *et al.* [11] | 93.2 |
| Text | | 87.1 |
| multimodal | | 95.8 |
| Image | Ferrando *et al.* [48] | 94.04 |
| Text | | - |
| multimodal | | 94.90 |
| Image | Cross-Modal [16] | 96.25 |
| Text | | **97.18** |
| multimodal | | **99.71** |
| Image | $\text{EAML}_{Tr-KLD_{Reg}}$ (Ours) | **97.99** |
| Text | | 96.27 |
| multimodal | | 98.57 |
| Baselines | Kumar *et al.* [90] | 43.8 |
| | Kang *et al.* [76] | 65.37 |
| | Afzal *et al.* [3] | 76.6 |
| | Harley *et al.* [61] | 79.9 |
| | Noce *et al.* [126] | 79.8 |

modalities (*i.e.* vision and language) -leading to great intra-dataset generalization- are an essential component for building generalized frameworks that lead to great inter-dataset generalization.

### Generalization Experiment Design on the Tobacco-3482 Dataset

To evaluate the generalization ability of our ensemble network trained on the RVL-CDIP dataset, we use the benchmark Tobacco-3482 dataset and report the overall accuracy, recall, precision, and F1-score as useful metrics to evaluate the performance of the single-modal and cross-modal modalities. Since the Tobacco-3482 is an imbalanced dataset, we focus more on the precision-recall metrics which are useful to measure the success of predictions when the classes are imbalanced, which are reported in Tables 3.5 and 3.6. Note that the precision metric is a measure of result relevancy, while the recall metric is a mea-

Table 3.5: The Inter-Dataset Evaluation results of the Mutual Learning $ML_{Tr-KLD_{Reg}}$ method on the Tobacco-3482 dataset.

| Mutual Learning ($ML_{Tr-KLD_{Reg}}$) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Class Labels | Vision Modality | | | Language Modality | | | Modality Fusion | | | #Nb. Samples |
| | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score | |
| Advertisement | 0.9659 | 0.9659 | 0.9103 | 0.9596 | 0.8261 | 0.8879 | 0.9772 | 0.9304 | 0.9532 | 230 |
| Email | 0.9688 | 0.9850 | 0.9768 | 0.9577 | 0.9833 | 0.9703 | 0.9673 | 0.9866 | 0.9769 | 599 |
| Form | 0.9484 | 0.8956 | 0.9212 | 0.9360 | 0.8817 | 0.9080 | 0.9408 | 0.9582 | 0.9494 | 431 |
| Letter | 0.8959 | 0.9718 | 0.9323 | 0.9035 | 0.9577 | 0.9298 | 0.9329 | 0.9806 | 0.9561 | 567 |
| Memo | 0.9562 | 0.9855 | 0.9706 | 0.9466 | 0.9726 | 0.9594 | 0.9717 | 0.9968 | 0.9841 | 620 |
| News article | 0.8650 | 0.9202 | 0.8918 | 0.8406 | 0.9255 | 0.8810 | 0.9146 | 0.9681 | 0.9406 | 188 |
| Resume | 0.9836 | 1 | 0.9917 | 0.9836 | 1 | 0.9917 | 0.9756 | 1 | 0.9877 | 120 |
| Scientific publication | 0.9462 | 0.3372 | 0.4972 | 0.8889 | 0.3372 | 0.4889 | 0.9368 | 0.3410 | 0.50 | 261 |
| Scientific report | 0.2907 | 0.2491 | 0.2683 | 0.2707 | 0.2340 | 0.2510 | 0.2773 | 0.2302 | 0.2515 | 265 |
| Overall Accuracy (%) | | | 84.82 | | | 83.72 | | | 86.68 | |

Table 3.6: The Inter-Dataset Evaluation results of the Ensemble Self-Attention Mutual Learning ($EAML_{Tr-KLD_{Reg}}$) approach on the Tobacco-3482 dataset.

| Ensemble Self-Attention Mutual Learning ($EAML_{Tr-KLD_{Reg}}$) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Class Labels | Vision Modality | | | Language Modality | | | Multimodal Fusion | | | #Nb. Samples |
| | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score | |
| Advertisement | 0.9910 | 0.9565 | 0.9735 | 0.9911 | 0.9696 | 0.9802 | 0.9865 | 0.9565 | 0.9713 | 230 |
| Email | 0.9916 | 0.99 | 0.9908 | 0.9933 | 0.99 | 0.9916 | 0.99 | 0.99 | 0.99 | 599 |
| Form | 0.9628 | 0.9606 | 0.9617 | 0.9630 | 0.9652 | 0.9641 | 0.9627 | 0.9582 | 0.9605 | 431 |
| Letter | 0.8983 | 0.9965 | 0.9448 | 0.9040 | 0.9965 | 0.9480 | 0.9056 | 0.9982 | 0.9497 | 567 |
| Memo | 0.9857 | 1 | 0.9928 | 0.9841 | 1 | 0.9920 | 0.9857 | 1 | 0.9928 | 620 |
| News article | 0.9490 | 0.9894 | 0.9688 | 0.9588 | 0.9894 | 0.9738 | 0.9487 | 0.9840 | 0.9661 | 188 |
| Resume | 0.9917 | 1 | 0.9959 | 0.9917 | 1 | 0.9959 | 0.9836 | 1 | 0.9917 | 120 |
| Scientific publication | 0.9519 | 0.3793 | 0.5425 | 0.9592 | 0.3602 | 0.5237 | 0.9364 | 0.3946 | 0.5553 | 261 |
| Scientific report | 0.2374 | 0.1774 | 0.2030 | 0.2261 | 0.1698 | 0.1940 | 0.2709 | 0.2075 | 0.2350 | 265 |
| Overall Accuracy (%) | | | 87.29 | | | 87.23 | | | 87.63 | |

sure of how many truly relevant results are returned. The F1-score measures the weighted average of the precision and recall, while the relative contribution of precision and recall to the F1-score are equal. However, we evaluate on 9 classes of the RVL-CDIP dataset which overlap with the classes of the Tobacco-3482 dataset, that are: Advertisement, Email, Form, Letter, Memo, News article, Resume, Scientific publication, and Scientific report. We exclude the category named Note from the Tobacco-3482 dataset which does not overlap with any of the categories of the RVL-CDIP dataset.

As it can be seen from Tables 3.5 and 3.6 and Figures 3.7 and 3.9a, the proposed $EAML_{Tr-KLD_{Reg}}$ method displays a better generalization behavior than the $ML_{Tr-KLD_{Reg}}$ method Figures 3.8 and 3.9b over 8 categories that overlap with the RVL-CDIP dataset. The $EAML_{Tr-KLD_{Reg}}$ method performs better with an overall accuracy of 87.29% for the vision modality, 87.23% for the language modality, and 87.63% for the multimodal fusion

Table 3.7: The Inter-Dataset Evaluation results of the Ensemble Self-Attention Mutual Learning ($\text{EAML}_{Tr-KLD_{Reg}}$) approach on the RVL-CDIP dataset.

| | Ensemble Self-Attention Mutual Learning ($\text{EAML}_{Tr-KLD_{Reg}}$) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Class Labels | Vision Modality | | | Language Modality | | | Multimodal Fusion | | |
| | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| Advertisement | 0.8292 | 0.9337 | 0.8783 | 0.8702 | 0.7281 | 0.7929 | 0.9381 | 0.9769 | 0.9571 |
| Email | 0.9654 | 0.9799 | 0.9726 | 0.9820 | 0.9366 | 0.9588 | 0.9944 | 0.9964 | 0.9954 |
| Form | 0.7953 | 0.9126 | 0.8499 | 0.9289 | 0.8746 | 0.9009 | 0.9588 | 0.9846 | 0.9715 |
| Letter | 0.9763 | 0.8109 | 0.8859 | 0.9417 | 0.8816 | 0.9106 | 0.9970 | 0.9574 | 0.9768 |
| Memo | 0.9660 | 0.8874 | 0.9250 | 0.9630 | 0.8926 | 0.9265 | 0.9972 | 0.9729 | 0.9849 |
| News article | 0.9577 | 0.7579 | 0.8462 | 0.9574 | 0.8076 | 0.8762 | 0.9966 | 0.9197 | 0.9566 |
| Resume | 0.9811 | 0.8802 | 0.9279 | 0.9985 | 0.9718 | 0.9850 | 0.9998 | 0.9891 | 0.9944 |
| Scientific publication | 0.5298 | 0.8827 | 0.6622 | 0.5218 | 0.9268 | 0.6677 | 0.5203 | 0.9856 | 0.6810 |
| Scientific report | 0.1858 | 0.0565 | 0.0867 | 0.3246 | 0.0974 | 0.1498 | 0.2889 | 0.0197 | 0.0368 |
| Overall Accuracy (%) | | | 78.89 | | | 79.06 | | | 86.68 |

modality, compared to 84.82%, 83.72%, and 86.68% for the $\text{ML}_{Tr-KLD_{Reg}}$ method respectively. Regarding the Scientific publication category, the recall of the model considering the $\text{EAML}_{Tr-KLD_{Reg}}$ and $\text{ML}_{Tr-KLD_{Reg}}$ methods is very low. Amongst all the samples, the ability of the model to find the positive samples of the Scientific publication category is only at 37.93%, 36.02%, and 39.46% for the vision modality, the language modality and the multimodal fusion modality respectively for the $\text{EAML}_{Tr-KLD_{Reg}}$ method, while it is at 33.72%, 33.72%, and 34.10% for each modality respectively for the $\text{ML}_{Tr-KLD_{Reg}}$ method. The low recall for the two methods is due to the overlap between two categories that are Scientific publication and Scientific report.

After all, we see that for the two proposed $\text{EAML}_{Tr-KLD_{Reg}}$ and $\text{ML}_{Tr-KLD_{Reg}}$ methods, the model returns very few results compared to the intra-dataset evaluation, but most of its predicted labels are correct when compared to the training labels for the single-modal modalities, as well as for the multimodal fusion modality. Amongst all classes, the generalization ability of the model given the two methods is very poor regarding the class Scientific report, where the precision and recall are very low, whereas, for the intra-dataset evaluation, the performance of the ensemble network concerning the category Scientific report is at 94.62%, and 94.30% for the multimodal fusion modality of the $\text{EAML}_{Tr-KLD_{Reg}}$ and $\text{ML}_{Tr-KLD_{Reg}}$ methods respectively.

Therefore, Table 3.8 illustrates the average-precision scores (AP) of the common categories for the two proposed methods $\text{ML}_{Tr-KLD_{Reg}}$, and $\text{EAML}_{Tr-KLD_{Reg}}$. Hence, we relate

Table 3.8: The average precision (AP) scores of the inter-dataset evaluation of the $\text{ML}_{Tr-KLD_{Reg}}$ and the $\text{EAML}_{Tr-KLD_{Reg}}$ for the multimodal Fusion modality on the Tobacco-3482 dataset.

| | Method | |
|---|---|---|
| Class Labels | $\text{ML}_{Tr-KLD_{Reg}}$ | $\text{EAML}_{Tr-KLD_{Reg}}$ |
| Advertisement | 0.94 | 1.00 |
| Email | 0.99 | 1.00 |
| Form | 0.97 | 0.99 |
| Letter | 0.98 | 0.99 |
| Memo | 0.99 | 1.00 |
| News article | 0.96 | 1.00 |
| Resume | 1.00 | 1.00 |
| Scientific publication | 0.50 | 0.69 |
| Scientific report | 0.28 | 0.29 |
| Micro-Average Precision | 0.86 | 0.91 |

a good generalization ability of our proposed $\text{EAML}_{Tr-KLD_{Reg}}$ and $\text{ML}_{Tr-KLD_{Reg}}$ methods trained on RVL-CDIP, and evaluated on Tobacco-3482, regarding 7 common classes between the RVL-CDIP and Tobacco-3482 datasets, except for the Scientific publication and the Scientific report categories where it generalizes the worst.

We illustrate in Figure 3.6 the precision-recall curves of the best and worst classes for the cross-modal modalities of the $\text{EAML}_{Tr-KLD_{Reg}}$ and $\text{ML}_{Tr-KLD_{Reg}}$ methods respectively. It shows the trade-off between precision and recall for different thresholds. We compute the average precision (AP) from prediction scores which summarizes a precision-recall curve. We see that the model is returning accurate results (high precision), as well as a majority of positive results (high recall), as it is the case for the categories Resume, Email, and Memo, where most of the predicted samples are labeled correctly for either the $\text{EAML}_{Tr-KLD_{Reg}}$ or the $\text{ML}_{Tr-KLD_{Reg}}$ methods. However, we observe a good precision but low recall for the Scientific publication category, and a bad precision and recall for the Scientific report category.

#### Generalization Experiment Design on the RVL-CDIP Dataset

Symmetrically, we propose to evaluate the generalization ability of our proposed model trained on the Tobacco-3482 dataset and validated on the large-scale RVL-CDIP dataset.

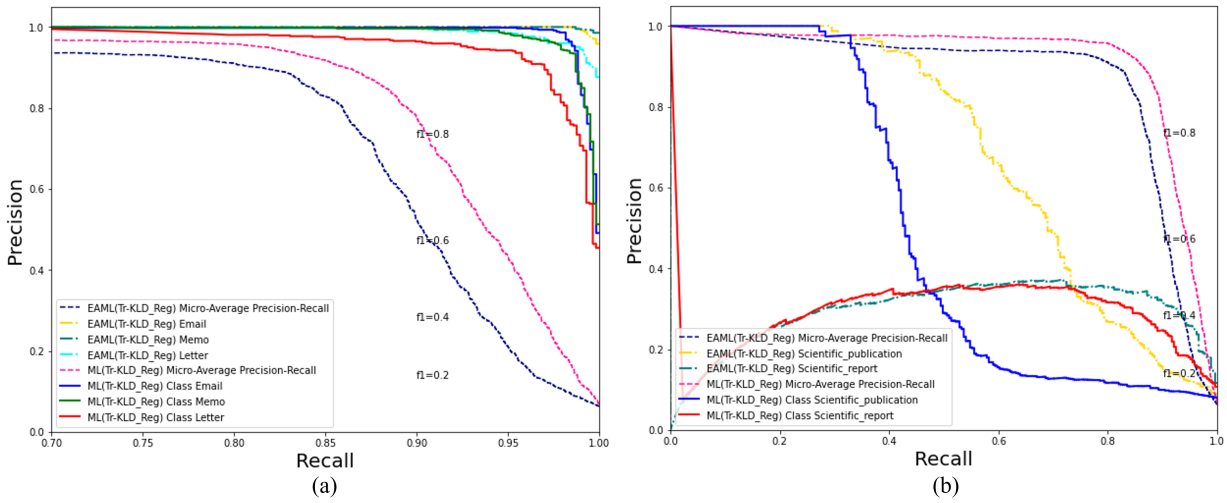



Figure 3.6: The Precision-Recall Curves of the Inter-Dataset Evaluation of the best and the worst classes of the cross-modal modalities for the two EAML$_{Tr-KLD_{Reg}}$ and ML$_{Tr-KLD_{Reg}}$ methods. (a) illustrates the P-R curves of the best classes. (b) illustrates the P-R curves of the worst classes.

The overall accuracy, recall, precision, and F1-score metrics of our best EAML$_{Tr-KLD_{Reg}}$ approach are proposed in Table 3.7. We proceed with the same evaluation protocol as in Section 3.5.4, where there are 9 classes of the Tobacco-3482 dataset that overlap with the classes of the RVL-CDIP dataset.

From Table 3.7, the EAML$_{Tr-KLD_{Reg}}$ method displays a better generalization ability compared to the other methods. It performs the best with an overall accuracy of 78.89% for the vision modality, 79.06% for the language modality, and 86.68% for the multimodal fusion modality. Amongst all classes, and similarly to the inter-dataset evaluation on the Tobacco-3482 dataset, the network generalizes the worst for the same categories which are Scientific publication and Scientific report, while it generalizes the best for the categories Resume, Letter, Memo, and Email. Moreover, the ensemble network manages to predict only 10.50% of samples that belong to the Scientific report category as true positives, while 85.26% are predicted as they belong to the Scientific publication category. At this stage, the precision and recall of the model are very low regarding the Scientific report category for each modality. As mentioned in Section 3.5.4, the bad precision and recall are due to the overlap between the two categories, which results to a bad generalization

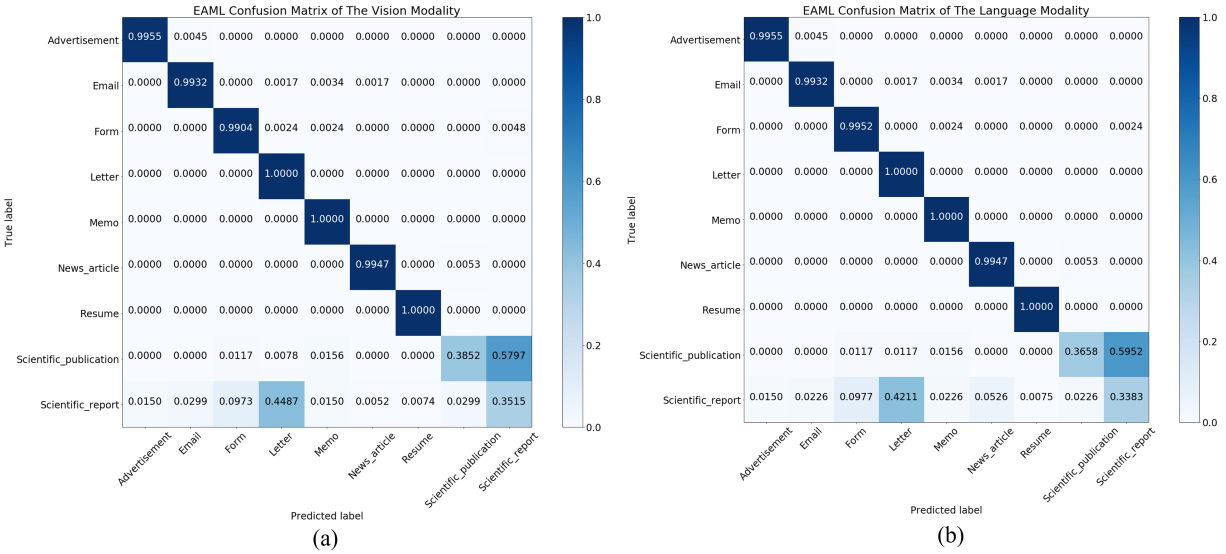



Figure 3.7: (a.) The Confusion Matrix of the Vision Modality of our best EAML$_{Tr-KLD_{Reg}}$) method. (b.) The Confusion Matrix of the Language Modality of our best EAML$_{Tr-KLD_{Reg}}$) method.

ability of the EAML$_{Tr-KLD_{Reg}}$ method considering only the two categories, contrary to the intra-dataset evaluation, where the ensemble network achieves accurate results with high precision and recall for all the categories.

Therefore, we relate a good generalization ability of our proposed EAML$_{Tr-KLD_{Reg}}$ trained on Tobacco-3482, and evaluated on RVL-CDIP, regarding 7 common classes between the RVL-CDIP and Tobacco-3482 datasets, except for the Scientific publication and the Scientific report categories where it generalizes the worst. These results are encouraging as we can see that our proposed system is able to learn on a smaller dataset consisting of $6k$ documents compared to the RVL-CDIP training set, which consists of $320k$.

### 3.5.5 Inter-Dataset Confusion Matrices

The Confusion matrices in Figure 3.7 display the generalization ability of our best EAML$_{Tr-KLD_{Reg}}$ approach for the vision, and language, modalities respectively. Symmetrically, Figure 3.8 refer to the vision, and language modalities of the proposed ML$_{Tr-KLD_{Reg}}$ method. Note that these methods are trained on RVLCDIP, and evaluated on the Tobacco-3482 dataset.

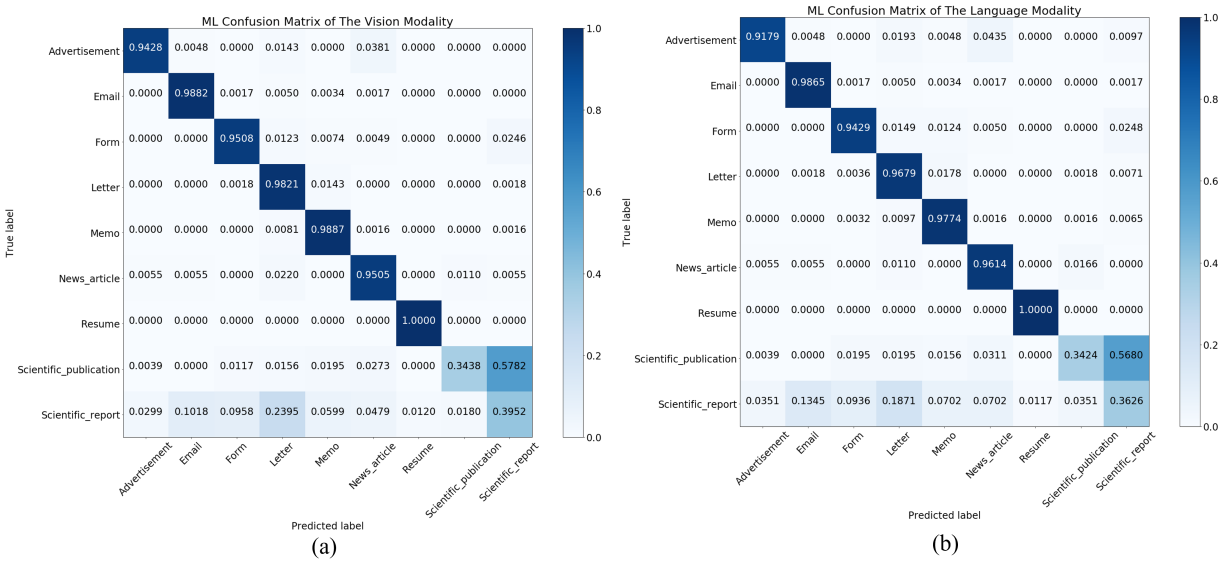



Figure 3.8: (a.) The confusion matrix of the Vision Modality of the $\text{ML}_{Tr-KLD_{Reg}}$ method. (b.) The confusion matrix of the Language Modality of the $\text{ML}_{Tr-KLD_{Reg}}$ method

## 3.6 Discussion

In this chapter, we have proposed a hybrid ensemble network that jointly learns the visual structural properties and the corresponding textual embeddings from document images through a self-attention-based mutual learning strategy ( *i.e.* $\text{EAML}_{Tr-KLD_{Reg}}$). We have shown that the designed self-attention-based fusion module along with the mutual learning approach with the regularization term enables the current modality to learn the positive knowledge from the other modality instead of the negative knowledge, which weakens the learning capacity of the current modality during the training stage. This constraint has been realized by adding a mimicry truncated-Kullback–Leibler divergence regularization loss (*i.e.* $\text{Tr-KLD}_{Reg}$) to the conventional supervised setting. With this approach, we have further combined the mutual predictions computed by the trained vision and language modalities in an ensemble network through multimodal learning to boost the overall classification accuracy of document images. The proposed mutual learning strategy with regularization has shown to be efficient in improving the overall performance of the ensemble model, as well the performance of the single-modal modalities. Finally, we displayed in detail the generalization capacity of our proposed models to classify unseen

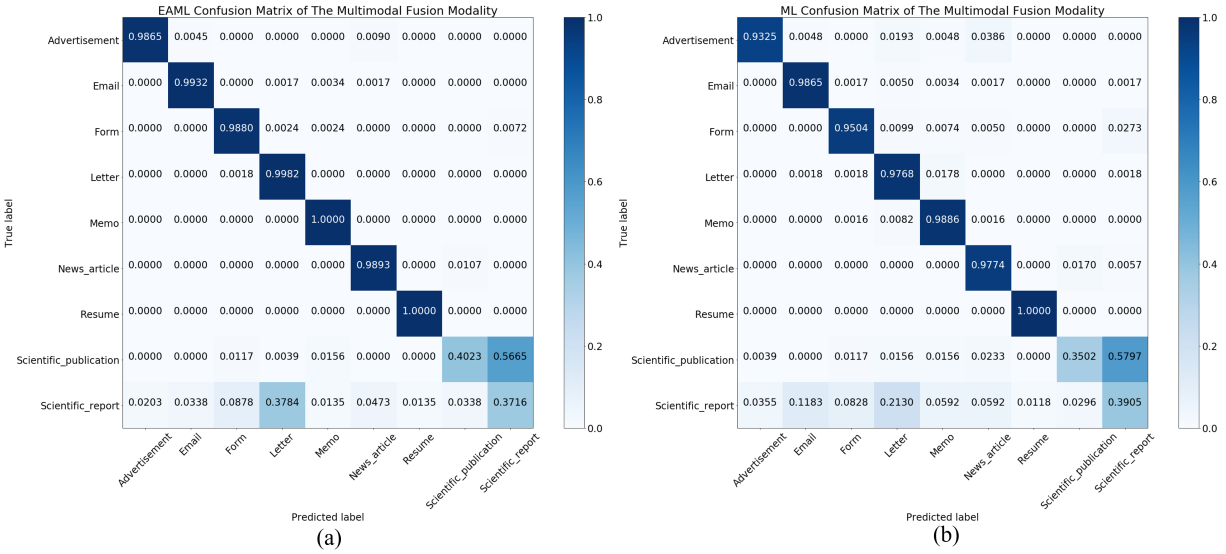



Figure 3.9: (a.) The confusion matrix of the Multimodal Fusion Modality of our best EAML$_{Tr-KLD_{Reg}}$ method. (b.) The confusion matrix of the Multimodal Fusion Modality of the ML$_{Tr-KLD_{Reg}}$ method.

document data by performing inter-dataset evaluation. We have demonstrated that cross-modal interactions and alignment between vision and language input queries in a fully-supervised manner are beneficial for developing a multimodal network that leads to good inter-dataset generalization.

Nevertheless, although these end-to-end multimodal deep neural networks often achieve superior performance, they have several limitations in real-world scenarios: (1) When performing cross-modal document classification during inference, the vision-language sample pairs need to be fed to the fusion modules to calculate the prediction scores to classify documents, which remains computationally expensive, as depicted in this Chapter as well as in Chapter 2. (2) To model high-level interactions between image regions and text sequences, in contrast to the previous related works that leverage different modalities into a joint embedding space, align them on the final embedding level, and thus, fail to model fine-grained interactions between the different modalities. In the Chapter 4, we address a more general and model-agnostic multimodal document understanding framework which is capable of learning more efficient cross-modal representations by modeling intra-modality and inter-modality interactions and relationships between vision and language modalities.

We make use of a cross-attention middle feature fusion transformer module to establish representation learning at the semantic level (by exploiting the relations between different document components). As well, specifically, we introduce a cross-modality learning strategy in the pre-training phase for contextualized comprehension on document components across vision and language modalities, which is further fine-tuned on two downstream applications which are document classification, and few-shot learning.

# CHAPTER 4

# Multimodal Document Representation Learning

> The deepest of level of communication is not communication, but communion. It is wordless... beyond speech... beyond concept.
>
> – *Thomas Merton*

## 4.1 Motivation

In Chapter 2, we developed a two-stream deep network to perform cross-modal document image classification based on an early feature fusion methodology (*i.e.* equal concatenation, average ensembling). As well, in chapter 3, we developed a multimodal deep network, trained in an end-to-end fully-supervised learning fashion, based on an intermediate self-attention feature fusion methodology. In contrast to the previous chapters, our goal is to develop a task-agnostic representation learning framework for document understanding in a pre-train-then-finetune paradigm. We aim to develop a domain-agnostic multimodal backbone for a better document understanding, by enhancing the cross-modal interactions within and across vision and language modalities. This leads to the following research questions of: **(1) can multimodal deep networks lead to task-agnostic cross-modal**

**representations for document data ?; (2) how to fully exploit visual and textual information of semantically meaningful components in document data, and to model the internal relationships among its components; (3) is a task-agnostic framework -pre-trained either on large-scale or low-scale document datasets- able to lead to domain-agnostic inter-dataset generalization over end-to-end fully-supervised learning frameworks, as established in Chapter 3 ?.**

Multimodal learning from document data has achieved great success lately as it allows us to pre-train semantically meaningful features as a prior into a learnable downstream approach. In this chapter we approach the document classification problem by learning cross-modal representations through language and vision cues, considering intra- and inter-modality relationships. Instead of merging features from different modalities into a common representation space, the proposed method exploits high-level interactions and learns relevant semantic information from effective attention flows within and across modalities. The proposed learning objective is devised between intra- and inter-modality alignment tasks, where the similarity distribution per task is computed by contracting positive sample pairs while simultaneously contrasting negative ones in the common feature representation spaces.

The recent research has started to consider how to leverage and incorporate the relations within those different modalities in a unified network to capture latent information for exploring better yet effective multimodal representations. Such systems have shown their effectiveness in improving multimodal representation learning in a pretrain-then-finetune paradigm, where models are first pre-trained with large-scale data and then fine-tuned to each downstream task [9, 21, 72, 101, 104, 139, 184, 185].

Several studies that have been devoted to perform the downstream document classification task, often used shallow cross-modal feature fusion modules to leverage visual and textual features such as naive concatenation, element-wise multiplication, and ensemble methods to extract cross-modal features [157, 188, 196]. Despite being studied extensively, the shortcomings of the preceding cross-modal feature fusion approaches are twofold. First, during inference, the vision-language sample pairs need to be fed to the fusion modules to calculate the prediction scores in order to perform the document classification task,

which remains computationally expensive. Second, the existing vision-language modality gap makes it difficult to capture high-level interactions between image regions and text sequences, as the feature representations of the visual and textual modalities are usually inconsistent and their distributions span different feature space.

In contrast, to embody the idea that better features make better classifiers, a framework that is based on the pretrain-then-finetune paradigm, which allows us to learn more general and task-agnostic cross-modal representations is highly required. Incorporating intra-modality and inter-modality relations from vision and language modalities can lead to more compact common representations. The resulting common representation space is an intermediate that implicitly measures the cross-modal similarities between image and text sequence sample pairs. Intuitively, the multimodality of documents requires multimodal reasoning over multimodal inputs, where data related to the same topic of interest tend to appear together. For instance, some types of documents such as handwriting categories are mainly not recognizable by OCR algorithms, which leads to losing textual information, and thus, semantic meaning. Thus, the visual information within the image regions of the document should be strongly emphasized. Meanwhile, some type of documents such as file folder category do not contain any visual spatial information, in which case a stronger emphasis on the textual information within the language cues is highly required.

To address the heterogeneity gap and the lack of closer interactions between image regions and text sequences within and across vision-language modalities, we propose a novel cross-modal contrastive pre-training model by learning cross-modal representations as a prior in a unified pre-training network. To encourage cross-modal learning, we model intra-modality and inter-modality representations between the cues of the vision-language modalities in the pre-training stage. We design an Inter-Modality Cross-Attention module denoted as (InterMCA) to capture relevant features from image regions and semantic meaning from text sequences. We aim to ensure that features from vision and language modalities map to closer points in the joint embedding space. Nevertheless, existing cross-modal document understanding approaches lack an explicit measure that also ensures that similar features from the same modality stay closer in the same joint embedding space. We assume that if similar features from the same category of each modality map

to distant points in the joint embedding space, then the embeddings generated within vision and language modalities will lack semantically enriched information, and thus, will generalize badly for downstream applications. As a remedy, we introduce intra-modality representation which is carried within an Intra-Modality Self-Attention module denoted as (IntraMSA). This module is devoted to constructing intra-modality relations within each modality according to the self-attention weights of image regions and text sequences.

Moreover, leveraging cross-modal relations through the InterMCA and IntraMSA attention modules requires a cross-modal learning objective. In the pre-training stage, we propose to train the network with a combinatorial cross-modal contrastive learning loss, which aims to simultaneously learn visual and textual features that represent document data in a more efficient manner than direct adoption of a single-modal contrastive loss for vision or language only modalities. For the downstream application, we run single-modal inference on top of the generated cross-modal embeddings to perform the specific document classification task. Also, we propose a new baseline in the few-shot setting. To the best of our knowledge, this is the first time to evaluate the generalization ability of a multimodal document embedding network on fewer samples in the document understanding field. The superior performance on the benchmark document datasets (*i.e.* RVL-CDIP and Tobacco-3482) demonstrates that the proposed cross-modal learning network, denoted as VLCDoC, can lead to learn robust and domain-agnostic cross-modal features in both document image classification and for few-shot document classification settings.

The main contributions of this work are summarized as follows:

- We design a unified task-agnostic document pre-training framework for a better cross-modal representation learning. Our network consists of leveraging two flexible extra levels of cross-modal interactions through cross-attention (InterMCA) and self-attention (IntraMSA) middle feature fusion-based attention modules. These modules capture high-level interactions between visual-textual cues within the different document components.

- We propose a cross-modal contrastive learning objective to further explore the relations between vision and language cues. Compared to the classic single-modal

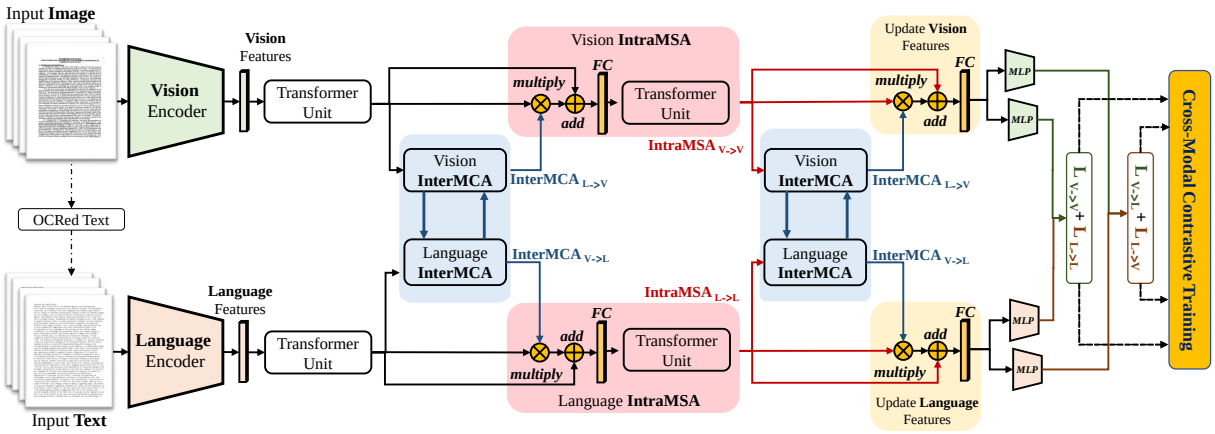



Figure 4.1: Overview of the proposed cross-modal contrastive learning method. The network is composed of InterMCA and IntraMSA modules with flexible attention mechanisms to learn cross-modal representations in a cross-modal contrastive learning fashion [19]

.

contrastive learning, the proposed cross-modal contrastive loss allows to learn and align the feature representations within and across vision-language modalities.

- Under a fair comparison setting, our task-agnostic framework demonstrates a good generalization ability among vision-language based approaches on the benchmark document datasets. It enables us to learn robust and domain-agnostic feature representations. Thus, it achieves better results compared to the generalization experiment design conducted in Chapter 3 for the document classification task.

- On the benchmark RVL-CDIP and Tobacco-3482 document datasets. We conduct for the first time in the document understanding literature, a new baseline on the few-shot learning setting. Thus, it achieves compelling results with significantly fewer document images used in the pre-training stage (*i.e.* when pre-trained on the Tobacco-3482 dataset).

## 4.2 Methodology

Figure 4.1 shows the overall architecture of the proposed cross-modal network. VLCDoC is an encoder-only transformer-based architecture trained in an end-to-end fashion. It

has two main modalities to perform visual-textual feature extraction. VLCDoC enforces deep multimodal interaction in transformer layers using a cross-modal attention module. The VLCDoC architecture network consists of two main schemes: one contrastive learning branch for cross-modal representation learning, and one cross-entropy learning branch for classifier learning. This feature learning strategy aims to learn a feature space which has the property of intra-class compactness and inter-class separability, while the classifier learning branch is expected to learn a domain-agnostic classifier with less bias based on the discriminative features obtained from the encoder branch.

### 4.2.1 Model Architecture

In this chapter, we design a multimodal transformer-based architecture for document understanding with unified cross-modal representation learning. Transformers have achieved great success in NLP, and are now heavily applied to images for different tasks such as image recognition, image classification, image captioning, image retrieval, and so on. Unlike deep CNNs which use pixel arrays, transformers applied to images (*i.e.* vision transformers (ViT)) split the images into visual tokens. The visual transformer divides an image into fixed-size patches, correctly embeds each of them, and includes positional embedding as an input to the transformer encoder. Moreover, ViT models have proven to be effective and outperform deep CNN models by almost four times when it comes to computational efficiency and accuracy.

#### Visual Features

To extract the visual embeddings, we follow the original pre-trained vision transformer architecture ViT-B/16 [46] as a backbone. Let $v_{visn} \in \mathbb{R}^{H \times W \times C}$ be the document image. We reshape it into a sequence of flattened $2D$ patches $v_{visn_p} \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where $(H, W)$ is the resolution of the document image, $C = 3$ is the number of channels, $(P, P)$ is the resolution of each document patch, and $N = HW/P^2$ is the resulting number of patches, which serve as the input sequence length for the transformer encoder. The patches obtained are then flattened and mapped to $d$ dimensions as the hidden embedding size. The resulting visual embeddings are then represented as $V = v^i_{visn} \in \mathbb{R}^{d_{visn}}$, where $d_{visn}$ is

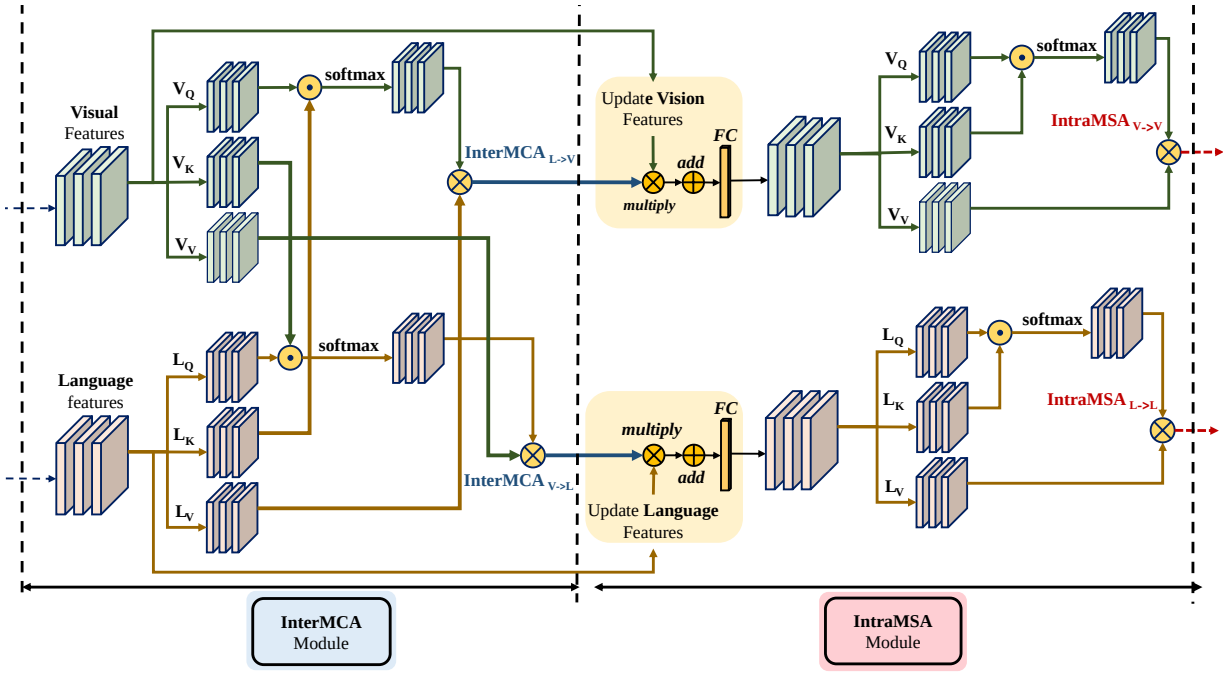



Figure 4.2: Illustration of the InterMCA and IntraMSA attention modules. The visual and textual features are transformed into query, key, and value vectors. They are jointly leveraged and are further fused to transfer attention flows between modalities to update the original features.

a $2D$ vector.

### Textual Features

To extract the textual embeddings, we first extract the text $t_{lang}$ within the document images via an off-the shelf optical character recognition (OCR) system, *e.g.* Tesseract OCR[1]. The input sequences extracted with the OCR are further fed into the pre-trained BERT$_{Base}$ uncased encoder [41]. The resulting textual embeddings are then represented as $T = t^i_{lang} \in \mathbb{R}^{d_{lang}}$, where $d_{lang}$ is a $2D$ vector of the same size as $d_{visn}$. This way, we ensure that the visual and the textual embeddings are of the same shape.

[1] https://github.com/tesseract-ocr/tesseract

### 4.2.2 Cross-Modal Alignment

In this subsection, we introduce the cross-attention (InterMCA) and self-attention (IntraMSA) modules that capture intrinsic patterns by modeling the inter-modality and intra-modality relationships for image regions and texts. Specifically, our proposed attention modules are transformer-based architectures as in [172]. It consists of a multi-head self-attention sub-layer, and a position-wise feed-forward sub-layer $f_{FF}$. Meanwhile, residual connections followed by the layer normalization $f_{LN}$ are also applied around each of the two sub-layers. In the multi-head self-attention sub-layer, the attention is calculated $h$ times, making it to be multi-headed. This is achieved by projecting the queries $\mathcal{Q}$, keys $\mathcal{K}$, and values $\mathcal{V}$ $h$ times by using different learnable linear projections.

#### Inter-Modality Alignment

The inter-modality cross-attention module InterMCA aims to enhance the cross-modal features by embracing cross-modal interactions across image regions and texts. This module aims to transfer the salient information from one modality to another as illustrated in Figure 4.2. Let $\mathbf{V}^l = \{v_1, v_2, ..., v_m\}$, $\mathbf{L}^l = \{l_1, l_2, ..., l_m\}$ be the sets of intermediate visual and textual features at the $l$-th layer of the vision and language modalities respectively, where $v_i \in \mathbb{R}^{1\times d_f}$, $l_i \in \mathbb{R}^{1\times d_f}$, and $\mathbf{V} \in \mathbb{R}^{m\times d_f}$, $\mathbf{L} \in \mathbb{R}^{m\times d_f}$. Note that the visual and textual features have the same dimensional feature vector $d_f$. To accomplish cross-modal interaction, we apply at first dot-product attention to combine the queries of each modality with the keys of the other. The weighted sum of the value of each modality is computed following the equations:

$$\textbf{InterMCA}_{\mathbf{L}\rightarrow\mathbf{V}}(\mathbf{V}^l) = \text{softmax}\left(\frac{\mathcal{Q}_{\mathbf{V}^l}\mathcal{K}^\top_{\mathbf{L}^l}}{\sqrt{d_k}}\right)\mathcal{V}_{\mathbf{L}^l} \tag{4.1}$$

$$\textbf{InterMCA}_{\mathbf{V}\rightarrow\mathbf{L}}(\mathbf{L}^l) = softmax\left(\frac{\mathcal{Q}_{\mathbf{L}^l}\mathcal{K}^\top_{\mathbf{V}^l}}{\sqrt{d_k}}\right)\mathcal{V}_{\mathbf{V}^l} \tag{4.2}$$

In this way, we emphasize the interaction and agreement between the visual regions and the semantic meaning of texts. The attention weights are then sent into the feed-forward

sub-layer. Finally, we get the output features of the next layer of the vision modality $\mathbf{V}^{l+1}$ computed as:

$$\mathbf{V}^{l}_{Att} = f_{LN_{\mathbf{V}}}(\mathbf{InterMCA}_{\mathbf{L}\rightarrow\mathbf{V}}(\mathbf{V}^{l}) + \mathbf{V}^{l}) \tag{4.3}$$

$$\mathbf{V}^{l+1} = f_{LN_{\mathbf{V}}}(f_{FF}(\mathbf{V}^{l}_{Att}) + \mathbf{V}^{l}_{Att}) \tag{4.4}$$

Similarly, the output features $\mathbf{L}^{l+1}$ of the language modality are computed as:

$$\mathbf{L}^{l}_{Att} = f_{LN_{\mathbf{L}}}(\mathbf{InterMCA}_{\mathbf{V}\rightarrow\mathbf{L}}(\mathbf{L}^{l}) + \mathbf{L}^{l}) \tag{4.5}$$

$$\mathbf{L}^{l+1} = f_{LN_{\mathbf{L}}}(f_{FF}(\mathbf{L}^{l}_{Att}) + \mathbf{L}^{l}_{Att}) \tag{4.6}$$

Further, the outputs of each vision and language InterMCA modules are subsequently fed into the vision and language IntraMSA modules.

### Intra-Modality Alignment

The IntraMSA attention module illustrated in Figure 4.2, aims to update the vision and language information and to capture inner-modality attention weights. For each modality, the information is updated according to a feature fusion scheme. At first, we perform element-wise product to the attention flow $\mathbf{V}^{l+1}$ with the the visual region features $\mathbf{V}^{l}$, then after a residual connection, features are fused by a linear additive function to yield the final updated visual information. To keep the dimension of the updated information consistent, a fully connected $f_{FC}$ layer is employed. The updated textual information is computed likewise, following the equations:

$$\hat{\mathbf{V}} = f_{FC}((\mathbf{V}^{l+1} \odot \mathbf{V}^{l}) + \mathbf{V}^{l}) \tag{4.7}$$

$$\hat{\mathbf{L}} = f_{FC}((\mathbf{L}^{l+1} \odot \mathbf{L}^{l}) + \mathbf{L}^{l}) \tag{4.8}$$

After updating original features based on cross-modal interactions, these features are fed into the transformer unit to intensify the inner-modality information, to preserve the original features and to establish inner-interactions simultaneously. Following the Equations 4.1 and 4.2, we have:

$$\mathbf{IntraMSA}_{\mathbf{V}\rightarrow\mathbf{V}} = \text{softmax}\left(\frac{\mathcal{Q}_{\hat{\mathbf{V}}^l}\mathcal{K}_{\hat{\mathbf{V}}^l}^{\top}}{\sqrt{d_k}}\right)\mathcal{V}_{\hat{\mathbf{V}}^l} \tag{4.9}$$

$$\mathbf{IntraMSA}_{\mathbf{L}\rightarrow\mathbf{L}} = \text{softmax}\left(\frac{\mathcal{Q}_{\hat{\mathbf{L}}^l}\mathcal{K}_{\hat{\mathbf{L}}^l}^{\top}}{\sqrt{d_k}}\right)\mathcal{V}_{\hat{\mathbf{L}}^l} \tag{4.10}$$

These two modules can be stacked repeatedly to enable the network to explore further latent intra-modality and inter-modality alignments between image regions and texts.

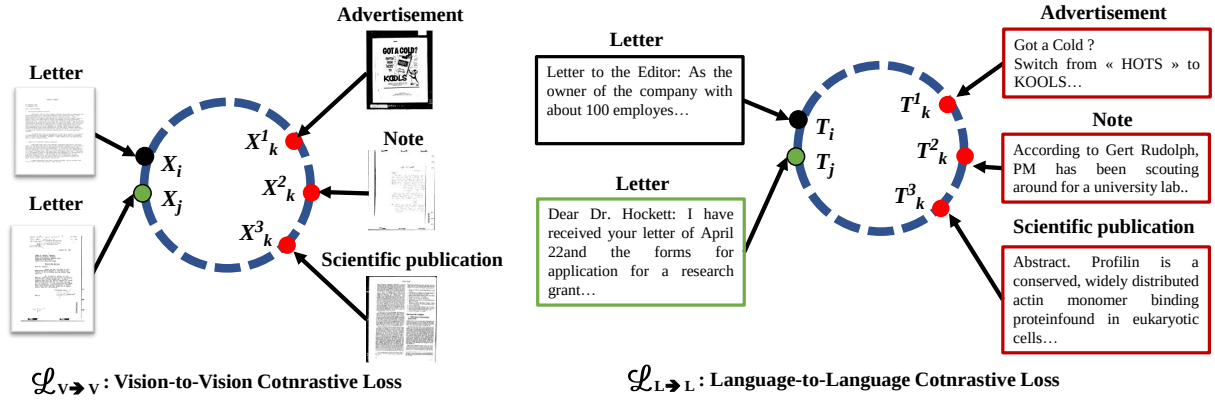
### 4.2.3 Cross-Modal Contrastive Learning

We design a visual-textual contrastive loss to force samples from language and vision that are semantically related to be closer.

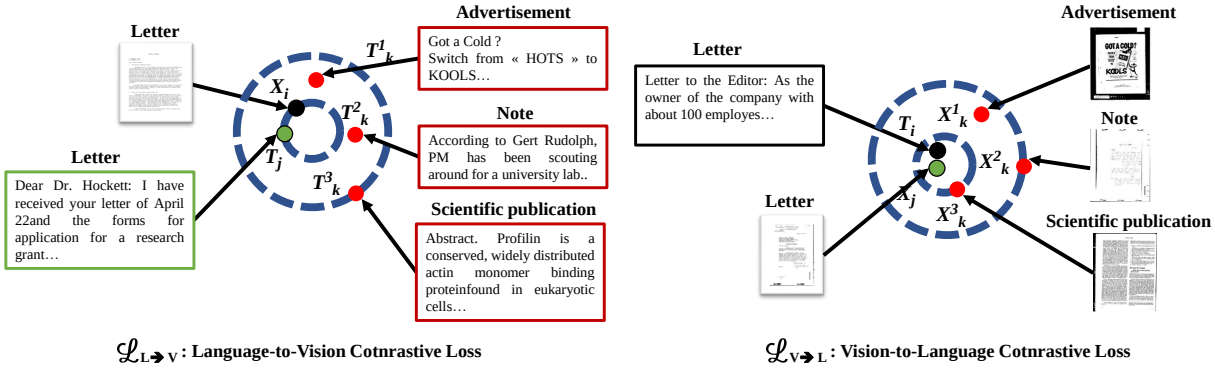
Besides, a projection head is implemented on top of the IntraMSA and InterMCA modules to map the image and text representations into a vector representation so that the two training schemes do not interfere with each other. The projection head is implemented as a nonlinear multiple-layer perceptron (MLP) with one hidden layer, as it is more suitable for contrastive learning [31]. Then, $L_2$ normalization is applied to the visual and textual embeddings so that the inner product between features can be used as distance measurements. In the following parts, we denote cross-modal contrastive learning as CrossCL.

#### Intra-Modality and Inter-Modality Contrastive Learning

Let $\{\mathbf{x}_i^+\} = \{x_j | y_j = y_i, i \neq j\}$, $\{t_i^+\} = \{t_j | y_j = y_i, i \neq j\}$ be the sets of all positive samples from the same class of an anchor image $x_i$ and an anchor text $t_i$ respectively, and $\{\mathbf{x}_i^-\} = \{x_j | y_j \neq y_i, \}$, $\{t_i^-\} = \{t_j | y_j \neq y_i\}$ be the sets of the remaining negative samples from other classes within the minibatch N. Not only should the pairs $(\mathbf{x}_i, \mathbf{x}_j)$, $(\mathbf{t}_i, \mathbf{t}_j)$ from




(a) Intra-modality contrastive learning




(b) Inter-modality contrastive learning

Figure 4.3: The proposed cross-modal contrastive learning objective

the same modality should be mapped to a close location in the joint embedding space (intra-modality), but also similar samples $\mathbf{x}_i$ and $\mathbf{t}_j$ should be mapped in close proximity (inter-modality). Therefore, the vision modality loss shown on the left of Figures 4.3a and 4.3b is computed as:

$$\mathcal{L}_V = \sum_{i=1}^{N} \mathcal{L}_{V \to V}(\mathbf{x}_i) + \sum_{i=1}^{N} \mathcal{L}_{L \to V}(\mathbf{x}_i) \tag{4.11}$$

$$\mathcal{L}_{V \to V}(\mathbf{x}_i) = \underbrace{\frac{-1}{|\{\mathbf{x}_i^+\}|} \sum_{\mathbf{x}_j \in \{\mathbf{x}_i^+\}} \log \frac{\exp(\mathbf{x}_i \cdot \mathbf{x}_j / \tau)}{\sum_{\mathbf{x}_k, k \neq i} \exp(\mathbf{x}_i \cdot \mathbf{x}_k / \tau)}}_{\text{Intra modality vision loss}} \tag{4.12}$$

$$\mathcal{L}_{L \to V}(\mathbf{x}_i) = \underbrace{\frac{-1}{|\{\mathbf{t}_i^+\}|} \sum_{\mathbf{t}_j \in \{\mathbf{t}_i^+\}} \log \frac{\exp(\mathbf{x}_i \cdot \mathbf{t}_j / \tau)}{\sum_{\mathbf{t}_k, k \neq i} \exp(\mathbf{x}_i \cdot \mathbf{t}_k / \tau)}}_{\text{Inter modality vision loss}} \tag{4.13}$$

where $\cdot$ computes similarity scores between example pairs and $\tau$ is a scalar temperature hyper-parameter. N is the minibatch size, $|\{\mathbf{x}_i^+\}|$ and $|\{\mathbf{t}_i^+\}|$ denote the number of positive samples of anchors $\mathbf{x}_i$ and $\mathbf{t}_i$ respectively. Similarly, the language modality loss shown on the right of Figures 4.3a and 4.3b is computed as:

$$\mathcal{L}_L = \sum_{i=1}^{N} \mathcal{L}_{L \to L}(\mathbf{t}_i) + \sum_{i=1}^{N} \mathcal{L}_{V \to L}(\mathbf{t}_i) \tag{4.14}$$

Therefore, the learning objective is based on four contrastive components including $V \to V$, $L \to V$, $L \to L$, and $V \to L$ alignments, which is computed as:

$$\mathcal{L}_{CrossCL} = \mathcal{L}_{V \to V} + \lambda \mathcal{L}_{L \to V} + \mathcal{L}_{L \to L} + \lambda \mathcal{L}_{V \to L} \tag{4.15}$$

where $\lambda$ is a hyper-parameter to control inter-modality alignment.

## 4.3 Experiments

In this section, we evaluate the effectiveness of the proposed method on low-scale and large-scale document classification datasets. We make use of the two benchmark datasets RVL-CDIP and Tobacco-3482 introduced in Section 1.5.1.

### 4.3.1 Pre-Training VLCDoC

The proposed VLCDoC method is implemented in Tensorflow with 4 NVIDIA GeForce 12Gb RTX 2080Ti GPU. For the vision modality, documents are resized into a fixed size of $(H, W) = (224, 224)$. The image region feature vector extracted by the ViT-B/16 backbone is of $d_{visn} = (197, 768)$. The final vision representation which is fed into the projection head is of dimension $d = 768$. As for the textual data, we tokenize the plain text $t_{lang}$ using a word-peace tokenizer to get $t_{tok}$. Each input sequence is expected to start with a $[CLS]$ token, and should end with a $[SEP]$ token. The $t_{tok}$ is then represented as: $t_{tok} = [CLS], t_{tok_1}, t_{tok_2}, ..., t_{tok_n}, [SEP]$, where $n = 197$ is the maximum sequence length. For each document, if $n > 197$, the input sequence is truncated so that it fits the desired length. For sequences that are shorter than $n < 197$, they are padded until they are $n = 197$ long. In the pre-training phase, the model is trained using AdamW optimizer with a learning rate of $2e-5$, linear warmup ratio to 0.1 and a linear decay. We set the batch size to 64 and we use the pre-trained weights of both ViT-B/16 and $\text{BERT}_{Base}$ uncased backbones. We conduct pre-training for 100 epochs for the RVL-CDIP and Tobacco datasets. We use the Adam [81] optimizer with learning rate of $5e-5$. For Tobacco-3482 dataset, we split the original sets to 80% for training, and 10% for validation and test. The temperature parameter $\tau$ is set to 0.1, and $\lambda$ is set to 0.5. Note that we didn't use any type of data augmentation during pre-training, and we kept the OCRed text as is without any pre- or post-processing. Note that the InterMCA and IntraMSA modules in our method are flexibly stacked two times to enhance the modeling of inter-modality and intra-modality relations during pre-training. We split the query, key, and value vectors of the visual features and textual features into four heads and concatenate the results in different sub-spaces.

### 4.3.2 Fine-tuning on Multimodal Tasks

**Task I: Document Image Classification.** The document image classification task aims to predict the category of visually rich document images. We conduct experiments on the RVL-CDIP and the Tobacco-3482 datasets. We take the encoder outputs on the

special tokens [LANG] and [VISN] from the last IntraMSA module module as holistic representations of the textual and visual inputs, which are used as the inputs to the vision and language classifiers. The whole fine-tuning takes 20 epochs with a batch size of 64 and a learning rate of $5e-5$ for both datasets.

**Task II: Few-Shot Document Image Classification.**

Given a pre-trained embedding network from stage one (*i.e.* pre-training), meta-testing is applied to the model with an episodic manner. A few-shot $K$-way multimodal document image classification task can be illustrated as a $K$-way $C$-shot problem. Given $C$ labelled samples for each unseen class, the model should fast adapt to them to classify novel classes. The entire test set can be presented by $D = \{[(v_1, y_N), ..., (v_N, y_Y)], [(l_1, y_N), ..., (l_N, y_Y)]\}$, where $N$ is the total number of classes in $D$, $v, l$ are the samples from the test set with label $y$. For a specific $K$-way $C$-shot meta-task $T$, $Y = \{yi|i = 1, ..., K\}$ denotes the class labels randomly chosen from dataset $D$. Samples from these classes are randomly chosen to form a Support set and a Query set: (a) the support set for task $T$ is denoted by $S$, which contains $CK$ samples ($K$-way $C$-shot); (b) the query set is $Q$ where $n$ is the number of samples selected for meta-testing.

During the meta-testing stage, the proposed model is tested to learn an embedding function to map all input image and text samples from the same class to a mean vector $c$ in a description space as a class descriptor for each class. For class $k$, it is represented by the centroid of embedding features of test samples and can be obtained as:

$$\mathcal{C}_k = \frac{1}{|S_k|} \sum_{(v_i, y_i) \in S} \mathcal{F}(v_i, l_i) \tag{4.16}$$

where $F(v_i, l_i)$ is the embedding function initialized by the pretext task, $S_k$ is the test samples labelled with class $k$. As a metric learning based method, we employ a distance function $d$ and produce a distribution over all classes given a query sample $q$ from the query set $Q$:

$$\mathcal{P}(y=k|q) = \frac{\exp\left(-d(f(q), c_k)\right)}{\sum_{k'}^{K} \exp(-d(f(q), c_{k'}))} \tag{4.17}$$

Euclidean distance is chosen as distance function $d$. As shown in Equation 4.17, the distribution is based on a softmax over the distance between the embedding of the samples (in the query set) and the class descriptors. The loss in the meta-testing stage can then read:

$$\mathcal{L}_{meta} = d(f(q), c_k) + \log \sum_{k'} d(f(q), c_{k'}) \tag{4.18}$$

In the meta-testing stage we average the results over 600 experiments as in [32]. In each experiment, we randomly sample 5 classes from novel classes, and in each class, we also pick $k$ instances for the support set and 15 for the query set. We conduct experiments on the most common setting in few-shot classification. 1-shot and 5-shot classification (*i.e.* 1 or 5 labeled instances are available from each novel class). We use the pre-trained (VLCDoC) network as the embedding network, and perform 5-way classification for only novel classes during the meta-testing stage.

### 4.3.3 Ablation Study

In this subsection, we conduct ablation studies to characterize our VLCDoC network on the low-scale Tobacco dataset. We analyze the following contributions of: i) validating the effectiveness of the proposed InterMCA and IntraMSA attention modules in learning generic cross-modal representations, ii) investigating whether contrastive learning enhances the cross-modal representations, resulting in a performance gain in terms of classification accuracy, iii) illustrating the generalization capacity and robustness of the proposed VLCDoC network.

Table 4.1: Ablation study on VLCDoC on cross-modality attention components, pre-trained on Tobacco dataset.

| Pre-training setting | IntraMSA | InterMCA | #Parameters | Accuracy(%) |
|---|---|---|---|---|
| *-w/o language modality* | | | | |
| | | | 198M | 85.71 |
| | √ | | 201M | 86.66 |
| | | √ | 209M | 87.20 |
| | √ | √ | 217M | **90.94** |
| *-w/o vision modality* | | | | |
| | | | 198M | 86.01 |
| | √ | | 201M | 86.31 |
| | | √ | 209M | 87.50 |
| | √ | √ | 217M | **90.62** |

**Effects of Attention Mechanisms**

To investigate the effectiveness of the attention mechanisms used in our VLCDoC model, we evaluate the performance of the learned cross-modal representations with and w/o the attention modules. Note that the evaluation protocol is single-modal based. At first, we consider the scheme where the vision and language modalities are pre-trained independently. In Table 4.1, we observe a significant drop to 85.71%, and 86.01% in classification performance when removing both attention mechanisms in the vision and language modalities respectively. When removing only the InterMCA module, we see that our model manages to improve slightly the performance of both modalities to 86.66% and 86.31% for the vision-language modalities. Note that at this stage, the pre-training of both modalities is still independent from one another. Further, removing the IntraMSA and keeping only the InterMCA module enables multimodal pre-training in an end-to-end fashion. The reported results in Table 4.1 show that our model gains in performance, and achieves the best performance with 90.94%, 90.62% top-1 accuracy for the vision and language modalities.

The improvement of the classification accuracy is attributed to the flexible attention flows adopted in both the InterMCA and IntraMSA modules, which have shown their effectiveness and capability to enhance vision-language relations by capturing the relevant

Table 4.2: Top-1 accuracy (%) comparison results of our proposed cross-modal contrastive learning loss against the standard supervised contrastive learning (SCL) loss on the Tobacco dataset.

| Model | Modality | CrossCL(%) | SCL(%) |
|---|---|---|---|
| **VLCDoC** | Vision | **90.94** | 89.88 |
| | Language | **90.62** | 89.29 |

semantic information of images and sentences. The results demonstrate the effectiveness of cross-modal learning and the importance of both attention modules in learning more effective cross-modal representations during the pre-training stage.

### Effects of Cross-Contrastive Learning

The Cross-modal Contrastive Loss (CrossCL) contains two components: the intra-modality alignment, and inter-modality alignment. We show the effects of cross-modal contrastive learning (CrossCL) on the proposed method against the standard supervised contrastive learning (SCL) loss. Table 4.2 shows that the CrossCL loss has a positive impact on the results. The VLCDoC with cross-modal contrastive learning loss CrossCL yields the best performance gain compared to VLCDoC with the Supervised Contrastive Loss (SCL). This indicates the importance of CrossCL by enforcing the compactness of intra-class representations (intra-modality), while separating inter-class features by contrasting positive and negative sample pairs within and across each modality. Note that, as described in Equation 4.15, the CrossCL can be vision cue-based or language cue-based, thus we have two different CrossCL presented in Table 4.2.

### Cross-Dataset Test

To illustrate the generalization capacity and the robustness of the learned cross-modal representations, we validate our proposed VLCDoC network on benchmark document classification datasets with different size and document types. We refer as the cross-dataset test to the process of pre-training our cross-modal network on dataset $A$, and fine-tune it and test it on dataset $B$. The motivation behind is to confirm whether our model displays a

Table 4.3: Cross-dataset test on datasets with different size and document types. Tobacco-3482, RVL-CDIP, Tobacco-3482 → RVL-CDIP denotes pre-train on the Tobacco-3482, fine-tune and test on RVL-CDIP.

| Model | Accuracy (%) | |
|---|---|---|
| | Tobacco-3482 → RVL-CDIP | RVL-CDIP → Tobacco-3482 |
| *w/o language modality* | | |
| - EAML [18] | 78.89 | 84.82 |
| - **VLCDoC** | **79.04** | **89.73** |
| *w/o vision modality* | | |
| - EAML [18] | 79.06 | 83.72 |
| - **VLCDoC** | **81.96** | **89.88** |

good generalization ability in terms of the downstream document classification task. Since there are no publicly available cross-document datasets for this specific task, we evaluate the ability of our model to perform document classification on a new set of documents that had not been seen by our model during the pre-training phase. For example, as denoted in Table 4.3, which refers to the cross-dataset test, RVL-CDIP→Tobacco denotes that the pre-training stage is firstly conducted on the RVL-CDIP dataset, then the fine-tuning stage of the previously pre-trained model is conducted on the Tobacco dataset. Finally, the test phase is conducted on the Tobacco dataset as well. Note that during the fine-tuning stage, we only train linear classifiers on the top of the final embeddings of the vision and language modalities of our pre-trained model, with the parameters of the rest of the layers freezed. Thus, even though the document categories are different between the dataset $A$ used for pre-training and test dataset $B$ used for fine-tuning and test, we can still evaluate our model on dataset $B$. The results confirm that our approach leads to a model with a better generalization ability compared to prior works.

As such, we compare our model with the related work EAML [18]. We first pre-train the model on the Tobacco dataset, then we conduct fine-tuning and test on the RVL-CDIP dataset. The reported results in Table 4.3 show that we slightly outperform EAML on both vision and language modalities. Even-though EAML is an ensemble network trained with a different setting, based on vision, language, and fusion modalities, the results confirm that our model benefits from cross-modal pre-training with a small amount of document

Table 4.4: Top-1 accuracy (%) comparison results of different document classification methods evaluated on the of RVL-CDIP dataset. V+L denotes vision+language modalities.

| Method | Pre-Training Data | Accuracy(%) | #Parameters |
|---|---|---|---|
| *vision methods* | | | |
| VGG-16 [4] | 320k | 90.31 | 138M |
| AlexNet [166] | 320k | 90.94 | 61M |
| ResNet-50 [4] | 320k | 91.13 | - |
| Ensemble [36] | 320k | 92.21 | - |
| DiT$_{Base}$ [101] | 320k | 92.11 | 87M |
| *(language+layout) methods* | | | |
| BERT$_{Base}$ [41] | - | 89.81 | 110M |
| RoBERTa$_{Base}$ [110] | - | 90.06 | 125M |
| LayoutLM$_{Base}$ [185] | 11M | 91.78 | 113M |
| *(vision+language) methods* | | | |
| w/o language | | | |
| - Multimodal [12] | 320k | 89.1 | - |
| - Ensemble [38] | 320k | 91.45 | - |
| - EAML [18] | 320k | 90.81 | - |
| w/o vision | | | |
| - Multimodal [12] | 320k | 74.6 | - |
| - Ensemble [38] | 320k | 82.23 | - |
| - EAML [18] | 320k | 88.80 | - |
| **VLCDoC (V+L)** w/o language | 320k | **92.64** | 217M |
| **VLCDoC (V+L)** w/o vision | 320k | **91.37** | 217M |
| *(vision+language+layout) methods* | | | |
| SelfDoc [104] | 320k | 93.81 | - |
| LayoutLM$_{Base}$ [185] | 11M | 94.42 | 160M |
| TILT$_{Base}$ [139] | 1M | 95.25 | 230M |
| LayoutLMv2$_{Base}$ [184] | 11M | 95.25 | 200M |
| LayoutLMv3$_{Base}$ [72] | 11M | 95.44 | 133M |
| DocFormer$_{Base}$ [9] | 5M | 96.17 | 183M |

data, achieving better performance with only vision and language modalities. Similarly, following similar protocol, we pre-train our encoder on RVL-CDIP, and then conduct fine-tuning and test on the Tobacco datasets with fewer document data. We clearly see that our model outperforms the work EAML with a significant margin of 4.91% and of 6.16% for vision and language modalities respectively. These results demonstrate that our model

displays a good generalization capacity which enables us to learn a robust and domain-agnostic feature representation for classifying documents with different document types and document data size.

**VLCDoC outperforms Baselines**

The comparison between the proposed VLCDoC network and existing methods on the large-scale RVL-CDIP document classification dataset is presented in Table 4.4. The compared methods cover various training strategies with different modalities used to perform document classification. These methods include (vision-only), (language-only), (vision-language), and (vision-language-layout) methods. Although our VLCDoC network learns feature space with vision and language cues, they use only single-modality (either vision or language) to classify document during the test. In Table 4.4, we can see that our VLCDoC model achieves the best performance with 92.64% and 91.37% of top-1 accuracy for using the vision or language modality respectively even compared to the methods that use the fusion of visual and language modalities. Note that the last group of methods use the layout as the supplementary information. For a fair comparison, we may integrate the layout information in our current framework as a new modality in the future work.

Therefore, the results reported demonstrate that our proposed approach outperforms all the methods that do not require any supplementary information such as layout information as used in [9, 72, 104, 184, 185]. Meanwhile, it achieves competitive results against the methods that include layout information in the pre-training setting. The results confirm that an encoder-only transformer-based architecture trained in an end-to-end fashion can help achieve compelling results against other methods which are mostly based on deep-CNN architectures.

**Few-Shot Classification.**

We also provide a scenario where fewer available instances can be accessed in document classification. To do so, we apply meta-testing on the test set of the RVL-CDIP and Tobacco-3482 datasets. We propose a new baseline in both intra-dataset and inter-dataset generalization by pre-training on dataset $A$, and testing on dataset $B$ as detailed in Ta-

Table 4.5: Intra-Dataset and Inter-dataset evaluation on the Few-shot document classification setting. The best embedding network is pre-trained on RVL-CDIP dataset, then tested on Tobacco-3482 dataset. All accuracy results are averaged over 600 test episodes and are reported with 95% confidence intervals.

| Pre-train Data | Fine-tune Data | Distance | Embedding Net | **1-Shot-5way** | **5-Shot-5way** | **20-Shot-5way** |
|---|---|---|---|---|---|---|
| RVL-CDIP | RVL-CDIP | Euclidean | VLCDoC (w/o Language) | 85.35 ± 0.046 % | 91.12 ± 0.015 % | 91.76 ± 0.015 % |
| | | | VLCDoC (w/o Vision) | 84.93 ± 0.046 % | 91.23 ± 0.015 % | 91.72 ± 0.015 % |
| | Tobacco-3482 | Euclidean | VLCDoC (w/o Language) | 54.31 ± 0.052 % | 66.61 ± 0.046 % | 71.81 ± 0.036 % |
| | | | VLCDoC (w/o Vision) | 53.29 ± 0.052 % | 66.40 ± 0.045 % | 72.16 ± 0.035 % |
| Tobacco-3482 | Tobacco-3482 | Euclidean | VLCDoC (w/o Language) | 48.00 ± 0.021 % | 54.22 ± 0.015 % | 61.24 ± 0.015 % |
| | | | VLCDoC (w/o Vision) | 47.32 ± 0.021 % | 56.33 ± 0.015 % | 61.51 ± 0.015 % |
| | Rvlcdip | Euclidean | VLCDoC (w/o Language) | 47.99 ± 0.021 % | 55.27 ± 0.018 % | 57.64 ± 0.017 % |
| | | | VLCDoC (w/o Vision) | 46.81 ± 0.021 % | 55.21 ± 0.018 % | 56.77 ± 0.016 % |

ble 4.5. We report the mean of 600 randomly generated test episodes as well as the 95% confidence intervals. On the common few-shot classification setting, the observation confirms that our multimodal document embedding network is also effective when pre-training and testing samples are rare across three tasks *i.e.* 1-shot, 5-shot, and 20-shot.

## 4.4 Discussion

In this chapter, we approached the document classification problem by proposing a novel cross-modal representation learning network, called VLCDoC, which models the intra-modality and inter-modality relations between visual and language cues via cross-modal contrastive learning. In addition, we introduced InterMCA and IntraMSA attention mechanisms by incorporating visual and textual features to further improve the cross-modal representations. A superior performance shows that a good generalization has been achieved with large-scale and low-scale datasets, which enables us to classify the document images in different domains. Although a compelling classification performance has not been achieved compared to related works that are based on vision+language+layout modalities, we aim to improve the multimodal representation learning of document images in the pre-training phase through self-supervised pretext-tasks. The general idea of Chapter 5 of this manuscript is to reduce the gap between vision and language-based methods and vision+language+layout based methods in terms of classification accuracy.

# CHAPTER 5

# Improved Multimodal Semantic Document Representation Learning

All our knowledge begins with the senses, proceeds then to the understanding, and ends with reason. There is nothing higher than reason.

– *Immanuel Kant*

## 5.1 Motivation

In the previous chapter, we addressed the problem of document image classification by learning cross-modal representations through contrastive learning by exploiting high-level interactions from effective attention flows within and across language and vision modalities. We have shown that the pre-trained embedding network is task-agnostic and enables us to generalize on fewer data in a domain-agnostic inter-dataset evaluation setting. In this chapter, our goal is to encourage multimodal interaction from language and vision in a self-supervised learning manner. Most multimodal pre-trained models rely on feature learning to learn their pretext objectives. Therefore, we propose to pre-train multimodal transformers with a two-step approach where feature learning and clustering are decoupled. We propose to develop a more general domain-agnostic and task-agnostic multimodal

document embedding model destined for document understanding applications. The idea of this chapter is to build a pre-trained multimodal document embedding, named LSRD (**L**earning Improved **S**emantic **R**epresentations for **D**ocument Understanding , which improves the semantic representation learning. At first, LSRD is pre-trained based on a nearest-neighbour instance discrimination technique to obtain semantically meaningful features. Second, we use the obtained features as a prior in a learnable clustering approach to remove the ability for cluster learning to depend on low-level features.

While most methods treat different views of the same image as positives for a contrastive loss, we are interested in using positives from other instances in the dataset. We propose to sample nearest neighbors from the dataset in the latent space, and treats them as positives in both vision and language modalities. This provides more semantic variations, as having more information helps in making more robust models. However, people learning from new data to be able to acquire newer concepts quickly depending on what they have already experienced is a key role in developing more complex multimodal machine learning algorithms having the same human subconscious understanding. Therefore, this assumption implies answering the following question of **how can an ability to find similarities across items of different modalities within previously seen samples improve self-supervised semantic representation learning ?**

With the recent rapid growth in the number of documents in business and academic fields, the annotation of large-scale documents is labor-intensive. Thus, learning multimodal knowledge from unlabeled documents is highly required, where the scale of the embedding network is constrained under a self-supervised learning (SSL) objective. Self-supervised pre-training techniques have been making remarkable progress recently in document representation learning [43, 55, 127, 182, 201]. Representation learning relies on pre-designed tasks, which do not require any annotated data to learn the weights of the multimodal embedding network. Instead, the multimodal features are learnt by minimizing the objective function of the pretext task. There have been extensive studies in the literature which include predicting the patch context in a given image [123, 182], inpainting patches [134], solving jigsaw puzzles [127, 129], colorizing images [94, 201], using adversarial training [44, 45], predicting noise [25], counting [128], predicting rotations [55],

spotting artifacts [75], generating images [145], using predictive coding [64, 130], performing instance discrimination [31, 62, 121, 168, 182], etc. Furthermore, recent studies in document representation learning rely on multimodal reasoning on multimodal input data (vision, language, and layout). For example, DocFormer [9] learns to reconstruct document image pixels through a CNN decoder, which tends to learn noisy details rather than high-level structures such as document layouts. SelfDoc [104] proposes to regress masked region features, which is noisier and harder to learn than classifying discrete features in a smaller vocabulary. LayoutLMv3 [72] learns to reconstruct masked word tokens of the language modality and symmetrically reconstruct masked patch tokens of the vision modality. Despite these efforts, representation learning approaches are mainly used as the first pre-training stage in a pre-train-then-finetune paradigm. The second stage includes fine-tuning the pre-trained network in a fully-supervised learning fashion on a specific downstream task, with the goal to verify how well the pre-trained embeddings transfer to the new downstream application.

Whereas most multimodal pre-trained document understanding techniques use Masked-Language Modeling (MLM) [41], Masked-Vision Modeling (MVM) [21], and Vision-Language Modeling (VLM) [72] techniques to learn cross-modal alignment between masked image patches and masked text tokens, we aim to study cross-modality learning for contextualized comprehension on document components across language and vision modalities in a self-supervised learning approach. We develop a model that learns both intra-modality and inter-modality relationships between visual and textual cues of document images through a multimodal attention feature fusion module following the architecture of Chapter 4. Given a collection of unlabelled documents, we attempt to learn a robust representation and maximize the mutual information between the vision and language modalities using nearest-neighbour contrastive learning as the self-supervision representation learning pretext task [31, 62, 182]. Further, after learning the feature representations, we propose to mine the vision-language nearest neighbours learnt through contrastive learning, based on a vision-language feature similarity approach, and use them as a prior into a learnable approach. We aim to classify each document image and its mined language content neighbours together by maximizing their dot product after applying the softmax objective

function. This strategy enables us to push the network to produce discriminative and consistent predictions. Thus, it prevents the cluster degeneracy scenario [28], leading to one cluster dominating the others by assigning all its probability mass to the same cluster when learning the decision boundary.

For downstream usage, we leverage the visual-textual semantic features learnt across our pre-trained embedding network into a feature fusion methodology for a more stable and better-performing solution for document-related downstream applications. One popular solution for few-shot classification is to apply a fine-tuning process on an existing embedding network to adapt to new classes. The main challenge is that the fine-tuning could easily lead to over-fitting, as only a few samples (1-shot, 5-shot, or 20-shot) for each class are available. One proposed solution for few-shot classification is a meta-learning process, in which the dataset is divided into subsets for different meta-tasks to learn how to adapt the model according to the task change. These methods highly rely on an effective pre-trained embedding network. As for content-based document retrieval, our goal is to evaluate the representation learning capability of our network to encode the input modalities in a meaningful way for cross-modal document retrieval. Retrieving data in documents generally relies on one modality (either vision or language). Thus, leveraging information from language and vision cues in an integrated fashion is crucial for developing an ideal system which proposes a diversity of ways in which document data could be used. We also investigate our model in the case where there is no annotation available.

The main challenge of our work is to design a pretext task which can exploit high-level compact visual-semantic representations that are useful for solving downstream tasks. The following are the main contributions in this chapter:

- We introduce multimodal nearest-neighbour contrastive learning to learn self-supervised representations that go beyond single instance positives as the first pretext task of our two-step pre-training approach.

- We propose to mine the multimodal nearest-neighbours learnt through contrastive learning as prior into a learnable approach, as the second pretext task, in order to produce consistent discriminative predictions.

- We evaluate our approach on cross-modal few-shot document classification, content-based document retrieval, and document classification. We show that our network can efficiently leverage the multimodal information from unlabeled documents which benefits from modeling the interaction between language and vision modalities in the model's pre-training stage.

- Experimental evaluation shows that our network outperforms prior works which are based on the vision-language modalities, and achieves compelling results compared to models which are based on vision, language, and layout modalities on the specific task of document classification.

- We address and explore two new downstream applications in document understanding, which are few-shot document classification and content-based document retrieval, to evaluate the effectiveness of the learnt multimodal representations to transfer to new tasks.

## 5.2 Method

In this section, we present the cornerstones of our approach. First, we show that instead of learning single instance positives (*i.e.* the instance discrimination task), multimodal nearest-neighbours are capable of learning better features that are invariant to the intra-class variability encountered in document images. To facilitate multimodal representation learning, we propose to pre-train multimodal transformers with unified vision-language objectives, following the architecture previously used in Chapter 4. LSRD learns more diverse positive pairs and thus better uni-modal representations before fusion using nearest-neighbor contrastive learning. Moreover, LSRD learns an alignment objective loss which predicts whether a pair of vision and language is matched (positive) or not matched (negative) after leveraging the visual-textual features into a joint feature-based transformer module. Second, we show how mining multimodal nearest-neighbors from the pretext task can be used as a prior into a learnable approach designed for semantic clustering. LSRD integrates the pre-trained multimodal features and learns a novel objective which

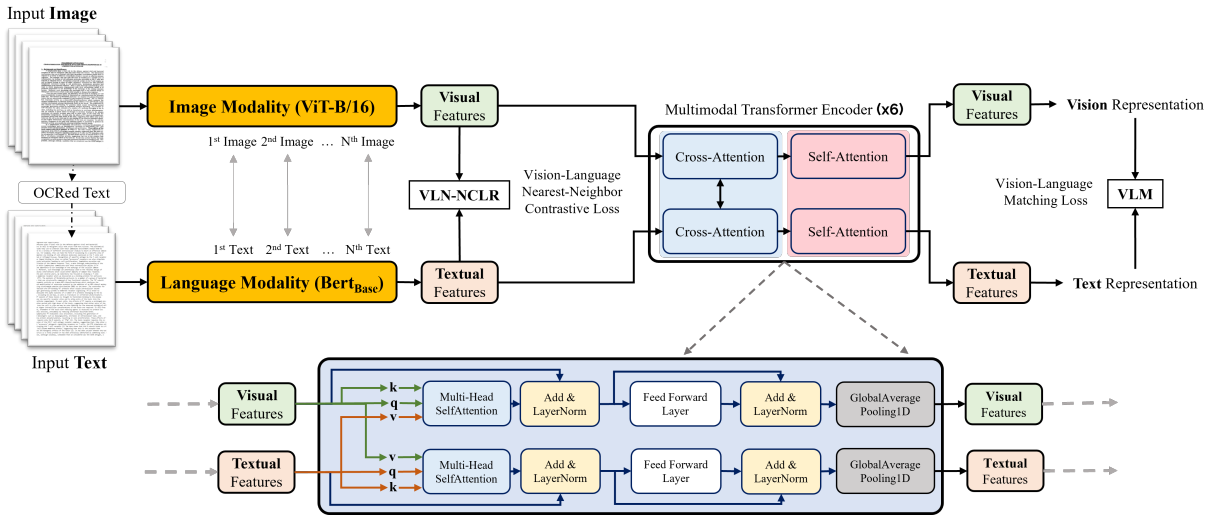



Figure 5.1: The architecture and pre-training representation learning objectives of LSRD. LSRD is a pre-trained multimodal transformer for document understanding with unified vision and language cross-modal learning objectives.

aims to classify each vision-language pair and their neighbours together.

### 5.2.1 Model Architecture

LSRD is an encoder transformer-based architecture. It applies a unified vision-language multimodal transformer to enforce deep multimodal interaction in transformer layers using novel multi-modal cross-attention feature fusion module. The multimodal transformer fusion network has a multi-layer architecture where each layer consists of multi-head self-attention and position-wise fully connected feed-forward networks [172]. Note that we use the same transformer architecture as in Chapter 4. The input of the multimodal transformer encoder is a concatenation of visual embeddings and language embeddings. Through the multimodal transformer, the last layer outputs vision-language contextual representations. Figure 5.1 illustrates the proposed LSRD approach.

#### Visual Features

Let $v_{visn} \in \mathbb{R}^{H \times W \times C}$ be the document image. We reshape it into a sequence of flattened $2D$ patches $v_{visn_p} \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where $(H, W)$ is the resolution of the document image, $C = 3$ is the number of channels, $(P, P)$ is the resolution of each document patch, and

$N = HW/P^2$ is the resulting number of patches, which serve as the input sequence length for the transformer encoder. The patches obtained are then flattened and mapped to $d$ dimensions as the hidden embedding size. The resulting visual embeddings are then represented as $V = v^i_{visn} \in \mathbb{R}^{d_{visn}}$, where $d_{visn}$ is a $1D$ vector.

#### Textual Features

To extract textual embeddings, we first extract the text $l_{lang}$ within document images via an off-the shelf optical character recognition (OCR) system, *e.g.* Tesseract OCR[1]. The input sequences extracted with the OCR are further fed into the pre-trained BERT$_{Base}$ uncased encoder [41]. The resulting textual embeddings are then represented as $L = t^i_{lang} \in \mathbb{R}^{d_{lang}}$, where $d_{lang}$ is a $1D$ vector of the same size as $d_{visn}$. This way, we ensure that the visual and the textual embeddings are of the same shape.

### 5.2.2 Pre-training Objectives

We pre-train LSRD with three objectives: vision-language nearest-neighbor contrastive learning (VLN-NCLR) on the uni-modal encoders, vision-language matching (VLM) on the multimodal encoder, and vision-language nearest-neighbor mining (VLN-NM) on the pre-trained embedding network.

**Objective I: Vision-Language Nearest-Neighbor Contrastive Learning (VLN-NCLR)**. We first describe the commonly used contrastive learning loss (*i.e.* InfoNCE) utilized in instance discrimination, and discuss NNCLR [47] which is based on nearest-neighbours of visual representations. Next, we introduce our approach, Vision-Language Nearest-Neighbor Contrastive Learning (VLN-NCLR) as cross-modal positives to learn better uni-modal representations before fusion, and thus, to improve contrastive instance discrimination between sample pairs from language and vision modalities.

**InfoNCE (Noise-Contrastive Estimation)** [130, 162, 182]: loss is commonly used in the instance discrimination setting [31, 62, 182]. The main idea is to pull representations of augmented versions/views of the same sample closer to each other (contracting positives),

[1] https://github.com/tesseract-ocr/tesseract

while simultaneously pushing different samples away from each other (contrasting negatives) in the representation space. For each given embedded sample $z_i$, another embedded sample (often a random augmentation of the sample) known as a positive sample pair $z_i^+$ is associated in addition to many negative embedding samples $z_i^- \in \mathcal{N}_i$. The InfoNCE loss is then defined as:

$$\mathcal{L}_i^{\text{InfoNCE}} = -\log \frac{\exp(z_i \cdot z_i^+/\tau)}{\exp(z_i \cdot z_i^+/\tau) + \sum_{z^- \in N_i} \exp(z_i \cdot z^-/\tau)} \tag{5.1}$$

where the sample pairs $(z_i, z_i^+)$ is considered as the positive pair, while $(z_i, z^-)$ is any negative pair in the minibatch. $\tau$ is the softmax temperature. In the vision-language context, given an image $x$ and its corresponding description $s$, we define the score function following Equation 5.1 as follows:

$$\mathcal{S}(v, l) = cos(f_{visn}(v), f_{lang}(l))/\tau \tag{5.2}$$

where $cos(v, l) = v^T l/||v||||l||$ denotes cosine similarity, and $\tau$ denotes a temperature hyper-parameter. $f_{visn}$ is an image encoder to extract the overall image feature vector and $f_{lang}$ is a text encoder to extract the global text feature vector. This maps the image and text representations into a joint embedding space $R^D$. The contrastive loss between image $v_i$ and its paired text $l_i$ is computed as:

$$\mathcal{L}v_i, l_i) = -log \frac{exp(cos(f_{visn}(v_i), f_{lang}(l_i))/\tau)}{\sum_{j=1}^{M} exp(cos(f_{visn}(v_i), f_{lang}(l_j))/\tau)} \tag{5.3}$$

The following two metrics are used for monitoring the pre-training performance:

(i.) **Contrastive accuracy**: Self-supervised metric, the ratio of cases in which the representation of an image is more similar to its corresponding text, than to the representation of any other image and text in the current batch.

(ii.) **Linear probing accuracy**: Linear probing is a popular metric to evaluate self-supervised classifiers. It is computed as the accuracy of a logistic regression classifier trained on top of the encoder's features. In our case, this is done by training a single

dense layer on top of the frozen encoder. Note that contrary to the traditional approach where the classifier is trained after the pre-training phase, in this example we train it during pre-training.

**NNCLR (Nearest-Neighbour Contrastive Learning)** [47]: proposes nearest-neighbours to obtain more diverse positive pairs by keeping a support set of embeddings which is representative of the full data distribution. To form the positive pairs, $z_i$'s nearest-neighbours are constructed from the support set $Q$. The NNCLR objective is then defined as:

$$\mathcal{L}_i^{\text{NNCLR}} = -\log \frac{\exp(\text{NN}(z_i, Q) \cdot z_i^+ / \tau)}{\sum_{k=1}^{n} \exp(\text{NN} \cdot z_k^+ / \tau)} \tag{5.4}$$

where NN(z, Q) is the nearest neighbour operator defined as:

$$\text{NN}(z, Q) = \arg\min_{q \in Q} || \, z - q \, ||_2 \tag{5.5}$$

**VLN-NCLR (Vision-Language Nearest-Neighbour Contrastive Learning):** The proposed learning objective VLN-NCLR aims to force samples from language and vision that are semantically related to be closer according to the computed nearest-neighbors of each modality. As in SimCLR [31], a projection head is implemented on top of the visual and textual embeddings to map the visual-textual representations into a vector representation so that the two training schemes do not interfere with each other. The projection head is implemented as a nonlinear multiple-layer perceptron (MLP) with one hidden layer, as it is more suitable for contrastive learning [31]. Then, $L_2$ normalization is applied to the visual-textual embeddings so that the inner product between features can be used as distance measurements. Building upon the NNCLR objective (see Equation 5.4), we define intra-modal and inter-modal learning losses defined as $\mathcal{L}_{Intra}^{\text{VLCLR}}$ and $\mathcal{L}_{Inter}^{\text{VLCLR}}$. For the intra-modal loss, it is composed of the vision modality loss $\mathcal{L}_{Visn \rightarrow Visn}^{\text{VLCLR}}$ and the language

modality loss $\mathcal{L}^{\text{VLCLR}}_{Lang \rightarrow Lang}$ which are computed respectively as:

$$\mathcal{L}^{\text{VLCLR}}_{Visn \rightarrow Visn} = \underbrace{-\log \frac{\exp(\text{NN}(v_i, \mathcal{V}) \cdot v_i^+ / \tau)}{\sum_{k=1}^{M} \exp(\text{NN}(v_i, \mathcal{V}) \cdot v_k^+ / \tau)}}_{\text{Intra-modality Vision loss}} \tag{5.6}$$

$$\mathcal{L}^{\text{VLCLR}}_{Lang \rightarrow Lang} = \underbrace{-\log \frac{\exp(\text{NN}(l_i, \mathcal{L}) \cdot l_i^+ / \tau)}{\sum_{k=1}^{M} \exp(\text{NN}(l_i, \mathcal{L}) \cdot l_k^+ / \tau)}}_{\text{Intra-modality Language loss}} \tag{5.7}$$

where $(\cdot)$ computes similarity scores between sample pairs and $\tau$ is a scalar temperature hyper-parameter, and M is the mini-batch size. Finally, the total intra-modal loss $\mathcal{L}^{\text{VLCLR}}_{Intra}$ can be written as:

$$\mathcal{L}^{\text{VLCLR}}_{Intra} = \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}^{\text{VLCLR}}_{Visn \rightarrow Visn}(v_i) + \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}^{\text{VLCLR}}_{Lang \rightarrow Lang}(l_i) \tag{5.8}$$

The second learning objective is an inter-modal loss which is composed of vision $\rightarrow$ language $\mathcal{L}^{\text{VLCLR}}_{Visn \rightarrow Lang}$ and language $\rightarrow$ vision $\mathcal{L}^{\text{VLCLR}}_{Lang \rightarrow Visn}$ losses. For the $\mathcal{L}^{\text{VLCLR}}_{Visn \rightarrow Lang}$ loss, it is computed as the similarity score between the nearest neighbors of the given document image $\text{NN}(v_i)$ and the corresponding text sample $l_i^+$. Similarly, the $\rightarrow$ vision $\mathcal{L}^{\text{VLCLR}}_{Lang \rightarrow Visn}$ loss is calculated as the similarity score between the nearest neighbors of the given text sample $\text{NN}(l_i)$ and its corresponding visual sample pair $v_i^+$:

$$\mathcal{L}^{\text{VLCLR}}_{Visn \rightarrow Lang} = \underbrace{-\log \frac{\exp(\text{NN}(v_i, \mathcal{V}) \cdot l_i^+ / \tau)}{\sum_{k=1}^{M} \exp(\text{NN}(v_i, \mathcal{V}) \cdot l_k^+ / \tau)}}_{\text{Inter-modality Vision loss}} \tag{5.9}$$

$$\mathcal{L}^{\text{VLCLR}}_{Lang \rightarrow Visn} = \underbrace{-\log \frac{\exp(\text{NN}(l_i, \mathcal{L}) \cdot v_i^+ / \tau)}{\sum_{k=1}^{M} \exp(\text{NN}(l_i, \mathcal{L}) \cdot v_k^+ / \tau)}}_{\text{Inter-modality Language loss}} \tag{5.10}$$

Finally, the inter-modal loss $\mathcal{L}^{\text{VLCLR}}_{Inter}$ is the sum of the vision and language losses over the mini-batch $M$.

$$\mathcal{L}^{\text{VLCLR}}_{Inter} = \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}^{\text{VLCLR}}_{Visn \rightarrow Lang}(v_i, l_i) + \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}^{\text{VLCLR}}_{Lang \rightarrow Visn}(l_i, v_i) \tag{5.11}$$

The NN($v_i$,$\mathcal{V}$), NN($l_i$,$\mathcal{L}$) denote the nearest neighbor operators, defined as:

$$\text{NN}(v_i, \mathcal{V}) = \arg\min_{q_v \in \mathcal{V}} || \, v_i - q_v \, ||_2 \tag{5.12}$$

$$\text{NN}(l_i, \mathcal{L}) = \arg\min_{q_l \in \mathcal{L}} || \, l_i - q_l \, ||_2 \tag{5.13}$$

**Objective II: Vision-Language Matching (VLM)**. The aim of this objective is to predict whether a pair of document images and their corresponding language is matched (positive) or negative (not matched). We compute the pairwise dot-product similarity between each language sequence $l_i$ and document image $v_i$ in the mini-batch as the predictions. The target similarity between the language sequence $l_i$ and the document image $v_i$ is computed as the average of the (dot-product similarity between $l_i$ and $l_j$) and (the dot-product similarity between $v_i$ and $v_j$). Then, the cross-entropy loss function is computed between the targets and the predictions. Given a mini-batch with $M$ document images and sequence samples, for each document image $v_i$, the vision-language pairs are constructed as $\{(v_i, l_j), y_{i,j}\}_{j=1}^{M}$, where $y_{i,j} = 1$ means that $(v_i, l_j)$ is a matched pair, while $y_{i,j} = 0$ indicates the unmatched ones. The probability of matching $v_i$ to $l_j$ is defined as:

$$\mathcal{P}_{i,j} = \frac{exp(v_i^T l_j)}{\sum_{k=1}^{M} exp(v_i^T l_k)} \tag{5.14}$$

where $l_j$ denotes the language feature vector, and $\mathcal{P}_{i,j}$ is the percent of scalar projection $v_i, l_j$ among all pairs $\{(v_i, l_j)\}_{j=1}^{M}$ in the mini-batch $M$. Geometrically, $v_i^T l_j$ represents the scalar projection vision feature vector $v_i$ onto the language feature vector $l_j$. The more similar vision feature to the language feature vector, the larger the scalar projection would be. Figure 5.2 shows the geometrical explanation of the cross-modal vision-language projection. Note that the scalar projection could be negative if the two feature vectors lie in opposite directions in the representation space. Then, the matching loss of associating $v_i$ with correctly matched language samples is defined as:

$$\mathcal{L}_{Visn \rightarrow Lang}^{\text{VLM}} = \frac{1}{M} \sum_{i=1}^{M} - \log(\mathcal{P}_{i,j}) \tag{5.15}$$

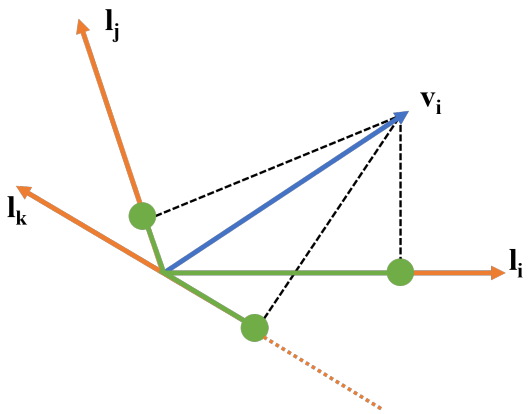



Figure 5.2: Interpretation of the cross-modal projection. The visual feature $v_i$ is projected onto different text directions $l_i, l_j and l_k$. The scalar projection of $v_i$ onto the matched text sequence $l_i$ is larger than that of unmatched text sequences $l_j$ and $l_k$.

In the vision-language matching scenario, the matching loss is usually computed in two directions as in [32, 111, 178]. The $Visn \rightarrow Lang$ matching loss requires the matched text to be closer to the document image than unmatched ones, and in verse the $Lang \rightarrow Visn$ matching loss constrains the related text to rank before unrelated ones. Similarly, the language matching loss $\mathcal{L}^{\text{VLM}}_{Lang \rightarrow Visn}$ can be formulated by exchanging $v$ and $l$ in Equation 5.14. Then, the total VLM loss is computed as follows:

$$\mathcal{L}^{\text{VLM}} = \mathcal{L}^{\text{VLM}}_{Visn \rightarrow Lang} + \mathcal{L}^{\text{VLM}}_{Lang \rightarrow Visn} \tag{5.16}$$

**Objective III: Vision-Language Nearest-Neighbor Mining (VLN-NM)** This objective aims to leverage the pretext features learnt across VLN-NCLR objective as a prior for clustering both the document images and their corresponding sequence samples. We motivated that a pretext task from representation learning can be used to obtain semantically meaningful features. Specifically, we freeze the multimodal embedding network obtained from representation learning pretext task (*i.e.* instance discrimination), and train only the last fully-connected layers on top of the pre-trained multimodal embedding network. For every document image sample $v_i \in M$ and its corresponding text sequences $l_i \in M$, we mine their $\mathcal{K}$ nearest neighbors in the embedding space $\Theta_\theta$. Let $\mathcal{N}_{v_i}$, and $\mathcal{N}_{l_i}$ be the sets of the neighboring samples of $v_i$ $l_i$ in the mini-batch $M$ respectively. We aim

to learn a clustering function $\Theta_\eta$ -parametrized by a deep neural network with weights $\eta$- that classifies a sample document image $v_i$, and a sample text sequence $l_i$ and their mined neighbors $\mathcal{N}_{v_i}$ $\mathcal{N}_{l_i}$ together. The function $\Theta_\eta$ terminates in a softmax function to perform a soft assignment over the vision clusters $\mathcal{C}_{Visn} = \{1, ..., C_{Visn}\}$ and language clusters $\mathcal{C}_{Lang} = \{1, ..., C_{Lang}\}$ with $\Theta_\eta(v_i) \in [0,1]^{C_{Visn}}$ and $\Theta_\eta(l_i) \in [0,1]^{C_{Lang}}$. The probabilities of sample pairs $v_i$, $l_i$ being assigned to cluster $C_{Visn}$, $C_{Lang}$ are denoted as $\Theta_\eta^{c_{Visn}}(v_i)$ and $\Theta_\eta^{c_{Lang}}(l_i)$ respectively. We then learn the weights of $\Theta_\eta$ by minimizing the following objectives for the vision modality:

$$\mathcal{L}_{Visn}^{\text{VLM}} = -\frac{1}{|M|}\sum_{v\in M}\sum_{k\in\mathcal{N}_v}\log\langle\Theta_\eta(v),\Theta_\eta(k)\rangle + \lambda\sum_{c_{Visn}\in\mathcal{C}_{Visn}}\Theta_\eta^{'c_{Visn}}\log\Theta_\eta^{'c_{Visn}} \tag{5.17}$$

$$\text{with;}\quad \Theta_\eta^{'c_{Visn}} = \frac{1}{|M|}\sum_{v\in M}\Theta_\eta^{c_{Visn}}(v) \tag{5.18}$$

with $\langle . \rangle$ denoting the dot product operator. The first term in Equation 5.17 forces $\Theta_\eta$ to make sure that neighbors have the same clustering assignment. Thus, to make consistent predictions for a sample document image or $v_i$ and its neighboring samples $\mathcal{N}_{v_i}$. Note that the dot product is maximal when the predictions are confident (one-hot) and assigned to the same cluster (consistent). In order to avoid $\Theta_\eta$ from assigning all samples to a single cluster, we include the second term in Equation 5.17, which is basically an entropy loss assigned to the clusters to make sure that the cluster distribution $\mathcal{C}$ is roughly uniform, so it can avoid assigning most of the document image instances to one cluster. Similarly to the vision modality, the language modality loss $\mathcal{L}_{Lang}^{\text{VLM}}$ can be written as:

$$\mathcal{L}_{Lang}^{\text{VLM}} = -\frac{1}{|M|}\sum_{l\in M}\sum_{k\in\mathcal{N}_l}\log\langle\Theta_\eta(l),\Theta_\eta(k)\rangle + \lambda\sum_{c_{Lang}\in\mathcal{C}_{Lang}}\Theta_\eta^{'c_{Lang}}\log\Theta_\eta^{'c_{Lang}} \tag{5.19}$$

$$\text{with;}\quad \Theta_\eta^{'c_{Lang}} = \frac{1}{|M|}\sum_{l\in M}\Theta_\eta^{c_{Lang}}(l) \tag{5.20}$$

In general, the number of clusters is unknown. However, similar to prior works [104], we choose $\mathcal{C}_{Visn}$ and $\mathcal{C}_{Lang}$ equal to the number of ground-truth clusters for the purpose of evaluation.

## 5.3 Experiments

### 5.3.1 Model Configurations

The proposed LSRD method is based on transformer encoders. For the vision modality, documents are resized into a fixed size of $(H, W) = (224, 224)$. The image region feature vector extracted by the $ViT - B/16$ backbone is of $d_{visn} = 768$. The final vision representation which is fed into the projection head is of dimension $d = 768$. As for the textual data, the textual feature vector is extracted by the $Bert_{BASE}$ as the language backbone. To pre-process the text input, we tokenize the plain text $t_{lang}$ using a Bert tokenizer to get $t_{tok}$. Each input sequence is expected to start with a $[CLS]$ token, and should end with a $[SEP]$ token. The $t_{tok}$ is then represented as: $t_{tok} = [CLS], t_{tok_1}, t_{tok_2}, ..., t_{tok_n}, [SEP]$, where $n = 256$ is the maximum sequence length. For each document, if $n > 256$, the input sequence is truncated so that it fits the desired length. Sequences that are shorter than $n < 256$ are padded until they are $n = 256$ long. We adopt distributed training and mixed-precision training to reduce memory costs and speed up training procedures. We also use a gradient accumulation mechanism to split the batch of samples into several mini- batches to overcome memory constraints for a large batch size. We adopt distributed training to reduce memory costs and speed up training procedures. We also use a gradient accumulation mechanism to split the batch of samples into several mini-batches to overcome memory constraints for a large batch size.

### 5.3.2 Pre-Training LSRD

In the pre-training phase, we use the training set of the RVL-CDIP document dataset to learn multimodal representations. LSRD is initialized from the pre-trained weights of the pre-trained vision and language backbones. For the multimodal transformer encoder, the weights are randomly initialized. We pre-train LSRD using the AdamW [112] optimizer with a batch size of 128 for $249,800$ steps. We use a weight decay of $1e-2$, $(\beta 1, \beta 2) = (0.9, 0.999)$. The learning-rate is warmed-up to $1e-4$ in the first 10% iterations, and decayed to $2e-5$ following a linear decay schedule. The temperature parameter $\tau$ is set

to 0.1, and the size of the queue used for vision-language contrastive learning is set as 65,536. Note that we didn't use any type of data augmentation during pre-training, and we kept the OCRed text as is without any post-processing.

### 5.3.3 Fine-Tuning on Multimodal Tasks

#### Task I: Document Image Classification

Table 5.1: Top-1 accuracy (%) comparison results of different document classification methods evaluated on the of RVL-CDIP dataset. V, T, and L denote Vision, Text, and Layout modalities.

| Method | Pre-Training Data | Modality | Accuracy(%) | #Params |
|---|---|---|---|---|
| CNN Ensemble [61] | 320k | V | 89.80 | *60M |
| VGG-16 [4] | 320k | V | 90.31 | 138M |
| GoogLeNet [35] | 320k | V | 90.70 | 13M |
| AlexNet [166] | 320k | V | 90.94 | 61M |
| Single Vision Model [166] | 320k | V | 91.11 | *140M |
| ResNet-50 [4] | 320k | V | 91.13 | - |
| Ensemble [36] | 320k | V | 92.21 | - |
| $\text{DiT}_{Base}$ [101] | 320k | V | 92.11 | 87M |
| LadderNet [152] | 320k | V | 92.77 | - |
| $\text{BERT}_{Base}$ [41] | - | T | 89.81 | 110M |
| $\text{RoBERTa}_{Base}$ [110] | - | T | 90.06 | 125M |
| $\text{LayoutLM}_{Base}$ [185] | 11M | T+L | 91.78 | 113M |
| $\text{LiLT}_{Base}$ [176] | 11M | T+L | 95.68 | 113M |
| VLCDoC [19] | 320k | V+T | 92.64 | 217M |
| SelfDoc [104] | 320k | V+T+L | 92.81 | - |
| **LSRD** | 320k | V+T | **93.19** | - |
| $\text{LayoutLM}_{Base}$ [185] | 11M | V+T+L | 94.42 | 160M |
| UDoc [60] | 11M | V+T+L | 95.05 | 272M |
| $\text{TILT}_{Base}$ [139] | 1M | V+T+L | 95.25 | 230M |
| $\text{LayoutLMv2}_{Base}$ [184] | 11M | V+T+L | 95.25 | 200M |
| $\text{LayoutLMv3}_{Base}$ [72] | 11M | V+T+L | 95.44 | 133M |
| $\text{DocFormer}_{Base}$ [9] | 5M | V+T+L | 96.17 | 183M |

The document image classification task aims to predict the category of visually rich document images. We conduct experiments on the RVL-CDIP dataset. We use pooled features to predict a classification label for a document. The whole fine-tuning takes 20

epochs with a batch size of 64 and a learning rate of $2e-5$. We report in Table 5.1 the classification performance on the test set, where the metric used is the Top-1 classification accuracy.

LSRD achieves state-of-the-art performance of 93.19% regarding the Vision+Text modalities. It outperforms our VLCDoC model introduced in Chapter 4. LSRD reduces the gap with related works based on three modalities (Vision+Text+Layout) which are pre-trained on much more training data (*i.e.* 11M) against 320k document images in our case.

### Task II: Few-Shot Document Image Classification

We conduct the same few-shot classification as in Chapter 4r. We use the pre-trained embedding network from stage one (*i.e.* pre-training), then apply meta-learning with an episodic manner. A few-shot $K$-way multimodal document image classification task can be illustrated as a $K$-way $C$-shot problem. Given $C$ labelled samples for each unseen class, the model should fast adapt to them to classify novel classes. The entire test set can be presented by $D = \{[(v_1, y_N), ..., (v_N, y_Y)], [(l_1, y_N), ..., (l_N, y_Y)]\}$, where $N$ is the total number of classes in $D$, $v, l$ are the samples from the test set with label $y$. For a specific $K$-way $C$-shot meta-task $T$, $Y = \{yi|i = 1, ..., K\}$ denotes the class labels randomly chosen from dataset $D$. Samples from these classes are randomly chosen to form a Support set and a Query set: (a) the support set for task $T$ is denoted by $S$, which contains $CK$ samples ($K$-way $C$-shot); (b) the query set is $Q$ where $n$ is the number of samples selected for meta-learning.

During the meta-learning stage, the proposed model is trained to learn an embedding function to map all input image and text samples from the same class to a mean vector $c$ in a description space as a class descriptor for each class. For class $k$, it is represented by the centroid of embedding features of test samples and can be obtained as:

$$\mathcal{C}_k = \frac{1}{|S_k|} \sum_{(v_i, y_i) \in S} \mathcal{F}(v_i, l_i) \tag{5.21}$$

where $F(v_i, l_i)$ is the embedding function initialized by the pretext task, $S_k$ is the test

Table 5.2: Few-shot classification accuracy results on the test set of the RVL-CDIP dataset. All accuracy results are averaged over 600 test episodes and are reported with 95% confidence intervals. MLP denotes the projector used on top of the vision and language modalities to perform (VLN-NCLR) pre-training objective. ME denotes the multimodal encoder used on top of the vision and language modalities to perform the vision-language matching pre-training objective (VLM). **('+', '-')** indicate results w/wo meta-learning.

| Pre-train Task | Inference Method | Projector | Modality | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Vision | | | Language | | |
| | | | **5-way/15-Query** | | | **5-way/15-Query** | | |
| | | | **1-shot** | **5-shot** | **20-shot** | **1-shot** | **5-shot** | **20-shot** |
| VLN-NCLR | LSRD$^-$ | wo/MLP | 34.66 ± 0.66 | 44.70 ± 0.64 | 51.15 ± 0.66 | 32.49 ± 0.63 | 41.73 ± 0.58 | 48.84 ± 0.54 |
| | LSRD$^-$ | w/MLP | **41.33 ± 0.71** | **61.55 ± 0.65** | **75.05 ± 0.49** | **38.02 ± 0.68** | **54.87 ± 0.66** | **68.92 ± 0.55** |
| | LSRD$^+$ | wo/MLP | 43.81 ± 0.71 | 63.63 ± 0.63 | 76.46 ± 0.51 | 38.91 ± 0.63 | 57.40 ± 0.61 | 72.57 ± 0.54 |
| | LSRD$^+$ | w/MLP | **53.51 ± 0.80** | **74.48 ± 0.67** | **82.86 ± 0.51** | **38.14 ± 0.61** | **57.04 ± 0.61** | **72.20 ± 0.56** |
| VLN-NCLR +VLM | LSRD$^-$ | w/ME | 54.89 ± 0.83 | 74.58 ± 0.62 | 82.91 ± 0.49 | 67.23 ± 0.96 | 77.82 ± 0.41 | 78.87 ± 0.35 |
| | LSRD$^+$ | w/ME | **67.23 ± 0.96** | **77.82 ± 0.41** | **78.87 ± 0.35** | **67.01 ± 0.94** | **77.53 ± 0.42** | **78.86 ± 0.35** |
| VLN-NCLR + VLM +VLN-NM | LSRD$^-$ | w/ME | 79.08 ± 0.88 | 89.10 ± 0.39 | 89.96 ± 0.37 | 75.45 ± 0.94 | 86.79 ± 0.41 | 88.45 ± 0.38 |
| | LSRD$^+$ | w/ME | **80.63 ± 0.64** | **89.36 ± 0.49** | **90.34 ± 0.38** | **79.77 ± 0.61** | **89.54 ± 0.56** | **90.33 ± 0.38** |

samples labelled with class $k$. As a metric learning based method, we employ a distance function $d$ and produce a distribution over all classes given a query sample $q$ from the query set $Q$:

$$\mathcal{P}(y = k|q) = \frac{\exp\left(-d(f(q), c_k)\right)}{\sum_{k'}^{K} \exp(-d(f(q), c_{k'}))} \tag{5.22}$$

Euclidean distance is chosen as distance function $d$. As shown in Equation 5.22, the distribution is based on a softmax over the distance between the embedding of the samples (in the query set) and the class descriptors. The loss in the meta-testing stage can then read:

$$\mathcal{L}_{meta} = d(f(q), c_k) + \log \sum_{k'} d(f(q), c_{k'}) \tag{5.23}$$

In contrast to Chapter 4, where we applied only the meta-testing stage to evaluate the ability of our pre-trained model to generalize on fewer data, in this chapter we introduce a novel baseline setting to perform both meta-training and meta-testing on top of the pre-trained multimodal embedding network to study the ability of our task-agnostic pre-trained multimodal embedding network to perform fine-tuning on fewer data in few-shot

setting. In the meta-training stage, the algorithm first randomly selects $N$ classes, and samples small base support set $S_b$ and a base query set $Q_b$ from document data samples within these classes. The objective is to train a classification model $\mathcal{M}$ that minimizes $N$-way prediction loss $\mathcal{L}_{N-way}$ of the samples in the query set $Q_b$. Here, the classifier $\mathcal{M}$ is conditioned on the provided support set $S_b$. By making the predictions conditioned on the given support set, a meta-learning method can learn how to learn from limited labeled data through training from a collection of tasks (*i.e.* episodes). In the meta-testing stage, all novel class data $X_n$ are considered as the support set for novel classes $S_n$, and the classification model $\mathcal{M}$ can be adapted to predict novel classes with the new support set $S_n$.

Different meta-learning methods have been applied in the literature to make predictions conditioned on the support set. We choose ProtoNet [161] as a first baseline to start with, which we denote as LSRD$^+$. In LSRD$^+$, the prediction of the samples in a query $Q$ is based on comparing the distance between the query feature and the support feature from each class as in Equation 5.22. LSRD$^+$ compares the euclidean distance between the query features and the class mean of the support features.

As detailed in Table 5.2, we evaluate the few-shot classification accuracy on the RVL-CDIP dataset. LSRD$^-$ denotes the results without meta-learning (*i.e.* only meta-testing is applied). We conduct experiments for each pre-training task (*i.e.* VLN-NCLR, VLN-NCLR+VLM, and VLN-NCLR+VLM+VLN-NM), with and without the projectors (*i.e.* MLP, and ME) For each task, the best-performing method is highlighted. We average the results over 600 experiments as in [32]. In each experiment, we randomly sample 5 classes from novel classes, and in each class, we also pick $k$ instances for the support set and 15 for the query set. We conduct experiments on the most common setting in few-shot classification: 1-shot, 5-shot, and 20-shot classification (*i.e.* 1 or 5 or 20 labeled instances are available from each novel class). We use the pre-trained LSRD network as the embedding network, and perform 5-way classification for only novel classes. During meta-training, we follow the data split strategy in [144] to sample document samples of 11 classes for fine-tuning, and 5 classes for testing. Note that we sample only 600 samples for each class. The results show that the two-step pre-training approach improves semantic

representation learning, and thus boosts the overall results of the vision-language modalities compared to a one-step only pre-training approach. Also, with a multimodal encoder, we learn better information by aligning and matching the image-text sample pairs. Furthermore, we observe that the performance of our proposed method significantly increases when receiving more samples as input (*i.e.* 20-shot) with/without meta-learning. To sum up, with the two-step pre-trained LSRD model, we demonstrate a good generalization ability when fine-tuned on fewer data. The experiments conducted in this work on the few-shot setting will be used as a baseline for future works to start with, as compelling performance has been achieved on both vision and language modalities.

**Task III: Uni-Modal and Cross-Modal document Retrieval**

To the best of our knowledge, this is the first time to evaluate the representation learning of multimodal document networks on the task of content-based retrieval. We focus on the evaluation of both uni-modal and cross-modal retrieval tasks to answer the question of **how useful are the multimodal representations encoded by the proposed LSRD task-agnostic model to solve queries in cross-modal retrieval tasks ?** Assuming the LSRD is already pre-trained, the problem of uni-modal and cross-modal document retrieval is then defined as follows: In the first phase, which corresponds to the indexing phase, we extract the vision and language backbones, and then, we generate the embeddings for all document images -in the dataset in which our model LSRD has been already pre-trained on- using the target modality only. In the second phase, which corresponds to the retrieval phase, we process the query modality using the pre-trained LSRD model without activating (*i.e.* with backbones frozen) the network of the target modality (*i.e.* which can be either vision or language). For example, let us carry out the task of vision $\rightarrow$ language retrieval, where we assume the query contains visual document data. The objective of this specific task is to retrieve relevant textual information contained in documents which belong to the same category as the given query visual document image. This is done by encoding all texts in the dataset using the pre-trained language backbone, and then, the query document image is sent using the pre-trained vision backbone. Further, we compare the embeddings of the query document image with the embeddings of the

Table 5.3: Quantitative evaluation results of Intra-Modal and Inter-Modal Content-based retrieval on RVLCDIP 40K test set in terms of Recall@K(R@K).

| Pre-train Tasks | Intra-Modal Retrieval | | | | | | Inter-Modal Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Vision → Vision | | | Language → Language | | | Vision → Language | | | Language → Vision | | |
| | **R@1** | **R@5** | **R@10** | **R@1** | **R@5** | **R@10** | **R@1** | **R@5** | **R@10** | **R@1** | **R@5** | **R@10** |
| VLN-NCLR | 78.85 | 91.21 | 94.23 | 74.52 | 90.00 | 93.67 | 5.37 | 14.60 | 21.20 | 4.73 | 13.30 | 19.29 |
| VLN-NCLR+VLM | 80.63 | 92.06 | 94.83 | 75.15 | 90.39 | 93.96 | 73.05 | 89.48 | 93.41 | 70.11 | 86.05 | 91.34 |
| VLN-NCLR+VLM+VLN-NM | **82.85** | **93.15** | **95.49** | **79.00** | **92.07** | **95.03** | **75.28** | **90.07** | **93.58** | **73.74** | **88.00** | **92.29** |

text data in a semantic way and retrieve the most similar ones. In this work, we evaluate the cross-modal retrieval tasks using the same pre-trained LSRD model. These tasks include Vision → Vision, Language → Language, Vision → Language, and Language → Vision. As an example, the Language → Vision retrieval corresponds to the task where the queries are texts and the retrieved samples are document images. As a performance measure of the ranking of the retrieved results, we use the Recall@K(R@K), which is a standard evaluation metric in content-based retrieval. We calculate the Recall@K(R@K) on a different number of samples to retrieve. As detailed in Table 5.3, we present results of different pre-training objectives in both intra- and inter-modal retrieval task. We obtain competitive results on the RVL-CDIP test set, which contains about 40k document images.

**Evaluation on VLN-NCLR Pre-training Task.** As reported in Table 5.3, we conduct the first experiments on content-based retrieval using the first pre-trained task (*i.e.* VLN-NCLR). We see that in the intra-modal retrieval setting, which corresponds to the unimodal retrieval. The pre-trained LSRD model achieves good performance in retrieving relevant information regarding the input query. Note that, given a document image as the vision query, we aim to retrieve the top-k relevant document images which belong to the same category as the query image. Similarly, given a query text as an input, we aim to retrieve the top-k relevant textual information that is contained in document images, which belong to the same category as the query text. Therefore, for the intra-modal retrieval task, we achieve good performance as the first new baseline in this chapter, with a better R@K score where we retrieve top-1, 5 top-5 and top-10 relevant document data with 78.85%, 91.21% and 94.23% accuracies respectively for the vision modality, and 80.63%, 92.02% and 94.83% accuracies respectively for the language modality. However,

for the inter-modal retrieval task, which corresponds to the cross-modal retrieval setting, the retrieval R@K score drops significantly for the two Vision → Language and Language → Vision tasks. This drop of R@K is mainly due to the fact that LSRD did not learn any cross-modal information and high-level interactions across vision and language modalities. Hence, the significant drop of the R@K scores for both modalities.

**Evaluation on VLN-NCLR+VLM Pre-training Task.** In this pre-training task, we add a multimodal transformer encoder to model the cross-modal interactions between vision and language modalities, with a matching learning objective. Here, we aim to overcome the problem of the first pre-training task when performed on the inter-modal (*i.e.* cross-modal) retrieval setting. Therefore, we can see from Table 5.3 that, with the multimodal encoder and the vision-language matching learning objective-by matching the vision-language sample pairs-, we do not only improve the scores of the R@K for the inter-modal setting with nearly 71.59% for the vision modality, and 70.06% for the language modality, given all R@K scores. Meanwhile, for the intra-modal setting, we boost the R@K scores with nearly 1.07% and 1.31% for the vision and language modalities respectively. Hence, the importance of learning high-level features with a multimodal transformer encoder in a matching learning objective.

**Evaluation on VLN-NCLR+VLM+VLN-NM Pre-training Task.** With this last pre-training task in our two-step approach (indicated as **('+') with meta-learning**, we aimed to improve the semantic representation learning of document data through a document semantic clustering approach. We highlight the reported results in Table 5.3. The results indicate that the best R@K scores have been achieved with the semantic clustering approach, which was performed on the pre-trained representation learning embedding (*i.e.* pre-trained with VLN-NCLR+VLM learning objectives). Therefore, we improve the R@K scores for all intra-modal (*i.e.* uni-modal) and inter-modal (*i.e.* cross-modal) retrieval tasks.

**Does a different sequence length of the query text help ?** So far we used a sequence length of 256 for downstream evaluation as in the pre-training stage. In Table 5.4, we vary the sequence length in the content-based retrieval task to see whether a larger or a smaller sequence is beneficial to retrieve more relevant information. We vary the sequence length

Table 5.4: Effects of sequence length on content-based document retrieval.

| Pre-training Tasks | Sequence length | Intra-Modal Retrieval Lang → Lang | | | Inter-Modal Retrieval Visn → Lang | | | Inter-Modal Retrieval Lang → Visn | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **R@1** | **R@5** | **R@10** | **R@1** | **R@5** | **R@10** | **R@1** | **R@5** | **R@10** |
| VLN-NCLR+VLM +VLN-NM | 512 | 71.57 | 89.48 | 93.82 | 71.06 | 88.51 | 92.88 | 68.79 | 85.43 | 90.92 |
| | 256 | **79.00** | **92.07** | **95.03** | **75.28** | **90.07** | **93.58** | **73.74** | **88.00** | **92.29** |
| | 128 | 73.17 | 89.12 | 93.18 | 71.09 | 88.51 | 92.87 | 68.39 | 84.55 | 89.80 |
| | 64 | 70.85 | 88.07 | 92.53 | 67.34 | 87.20 | 91.85 | 62.60 | 80.17 | 86.13 |
| | 32 | 62.99 | 84.88 | 90.32 | 62.27 | 85.02 | 90.64 | 52.76 | 71.20 | 77.18 |
| | 8 | 38.02 | 67.83 | 79.94 | 46.91 | 75.95 | 85.03 | 21.33 | 34.83 | 41.02 |

from $8, 32, 64, 128, 256, 512$. The reported results indicate that with the same sequence length of 256 as used in the pre-training stage, we manage to get the best R@K scores.

### 5.3.4 Qualitative Results

In this subsection, we show representative samples of the retrieval output of the pre-trained LSRD network on the test set of RVL-CDIP dataset. In each one of the Figures 5.3, 5.4, 5.5, 5.6, the first column corresponds to the input query, and the top 5 retrievals are shown in following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown in a red border. Retrievals from other classes are considered incorrect, but they are often good retrievals nonetheless.

#### Uni-Modal Document Retrieval

**Vision → Vision.** In Figure 5.3, it is interesting to notice that in the first row, in which the query document image is a form, the top seven retrievals are all different memos from the same author (with the same signature) as the memo in the query image. The final row is similarly impressive: every document in the top ten retrievals has the same letterhead as the query document, despite variations in the other content, and also despite differing typefaces of the letterhead. There may exist biases in the dataset that lead to such fortunate retrievals (e.g., only a few letterheads, and only a few memo authors), but the results are still remarkable.

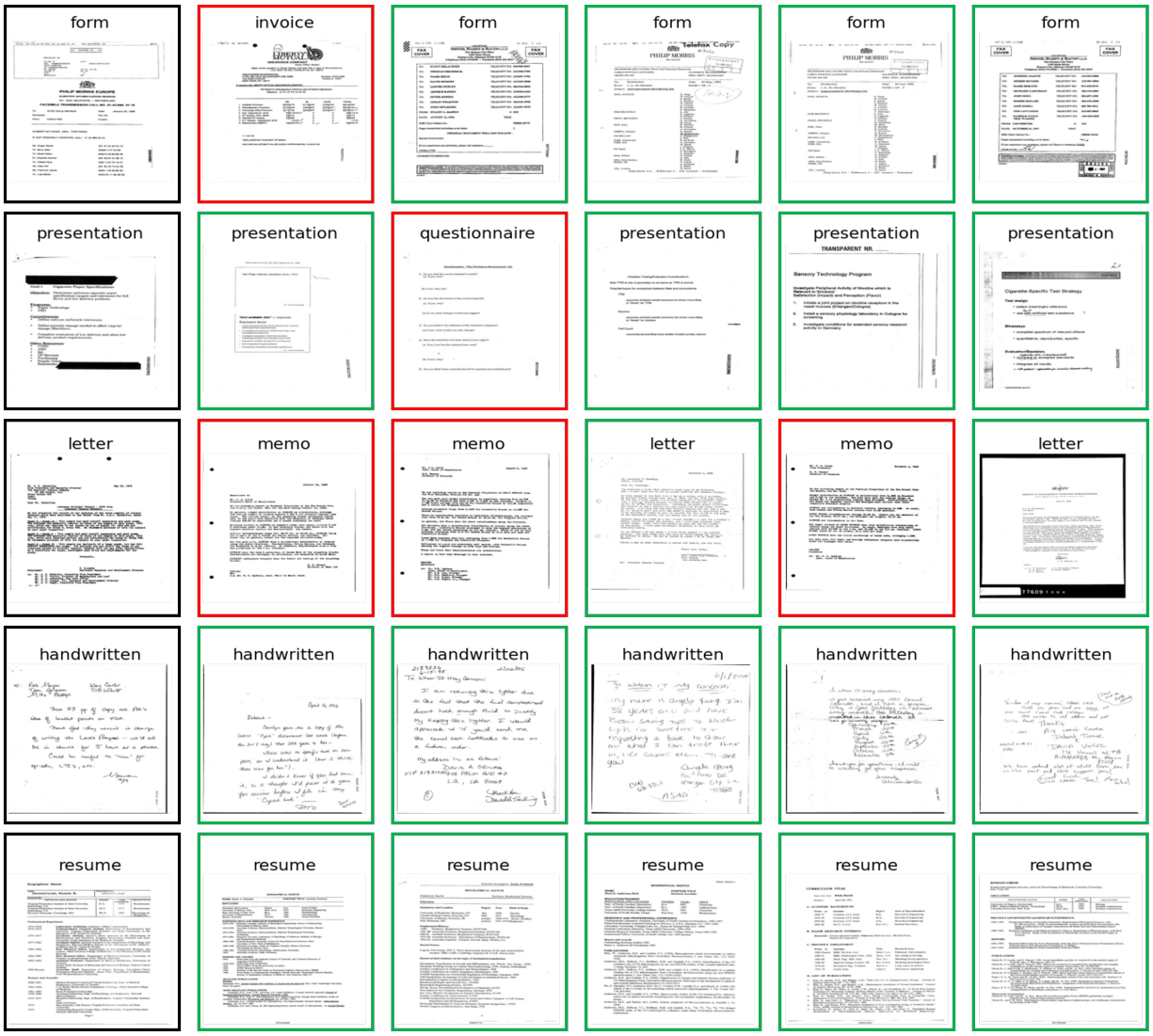



Figure 5.3: Vision to Vision Representative output of the retrieval process. Randomly selected Query document images are shown in the first column, and the top-5 document image retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border.

**Language → Language.** In Figure 5.4, the first row, in which the text query is a presentation, the top-5 text retrievals are all different from the query. The fourth row is similarly impressive: every document in the top-5 retrievals has the same letterhead as the query document, despite variations in the other content, and also despite differing typefaces of the letterhead. In spite of the differences in semantic meaning of the text

content between a form document and a letter document, the information shared from vision across the multimodal transformer encoder may have led to the incorrect prediction.

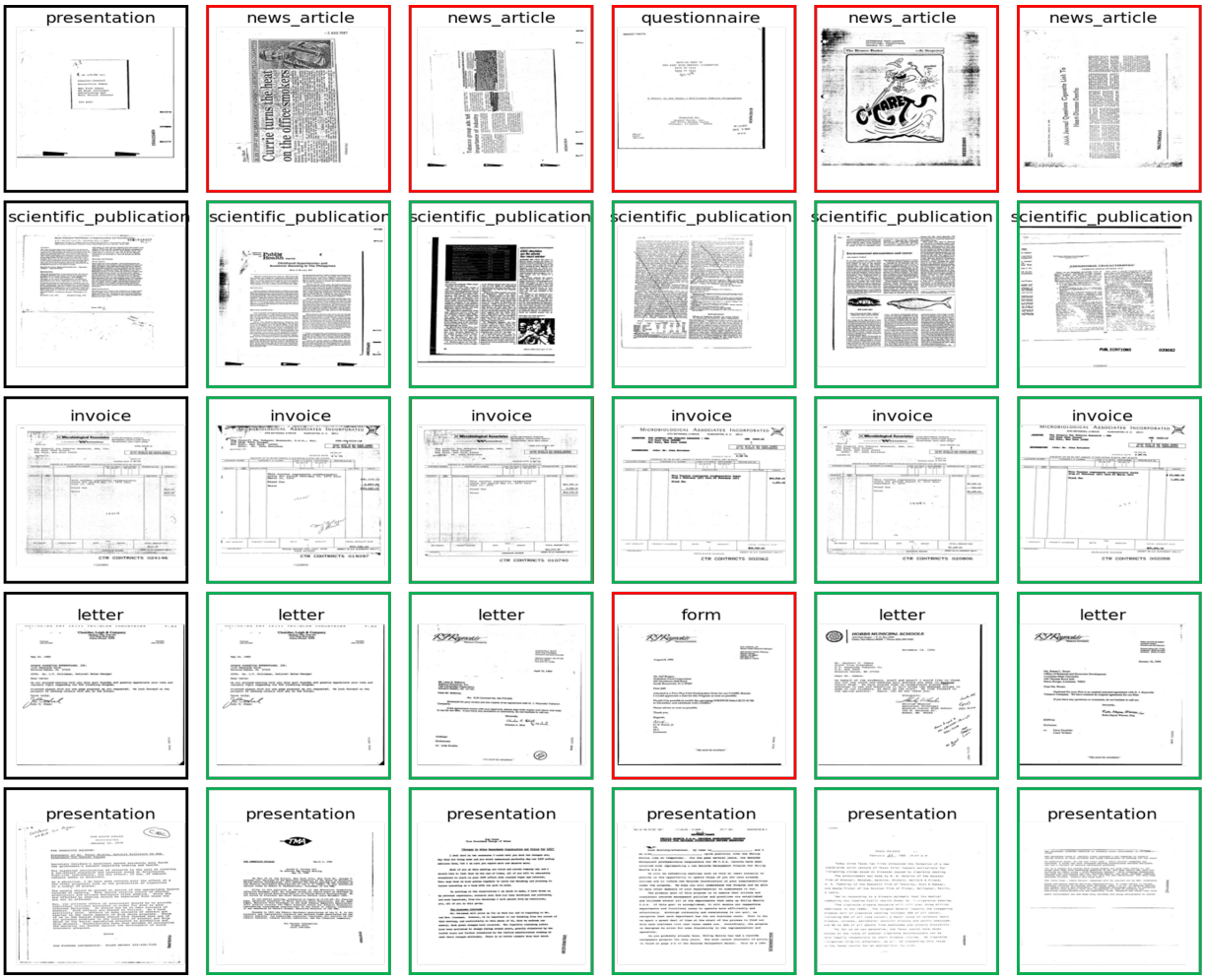



Figure 5.4: Language to Language Representative output of the retrieval process. Randomly selected Text sequences are used as query in the first column, and the top-5 text sequence retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the queries and retrieved results for a better visualisation.

### Cross-Modal Document Retrieval

**Vision → Language.** In Figure 5.5, we retrieve the top-5 relevant text content given a query document image. In the second row, we observe that the retrieved advertisement document is visually dissimilar compared to the query news_article document image.

Similarly, the top-5 retrieved text samples are all different to the query image in terms of the category label, and do not share the same visual information either. Therefore, the vision backbone is unable to retrieve similar text content of the given query document.

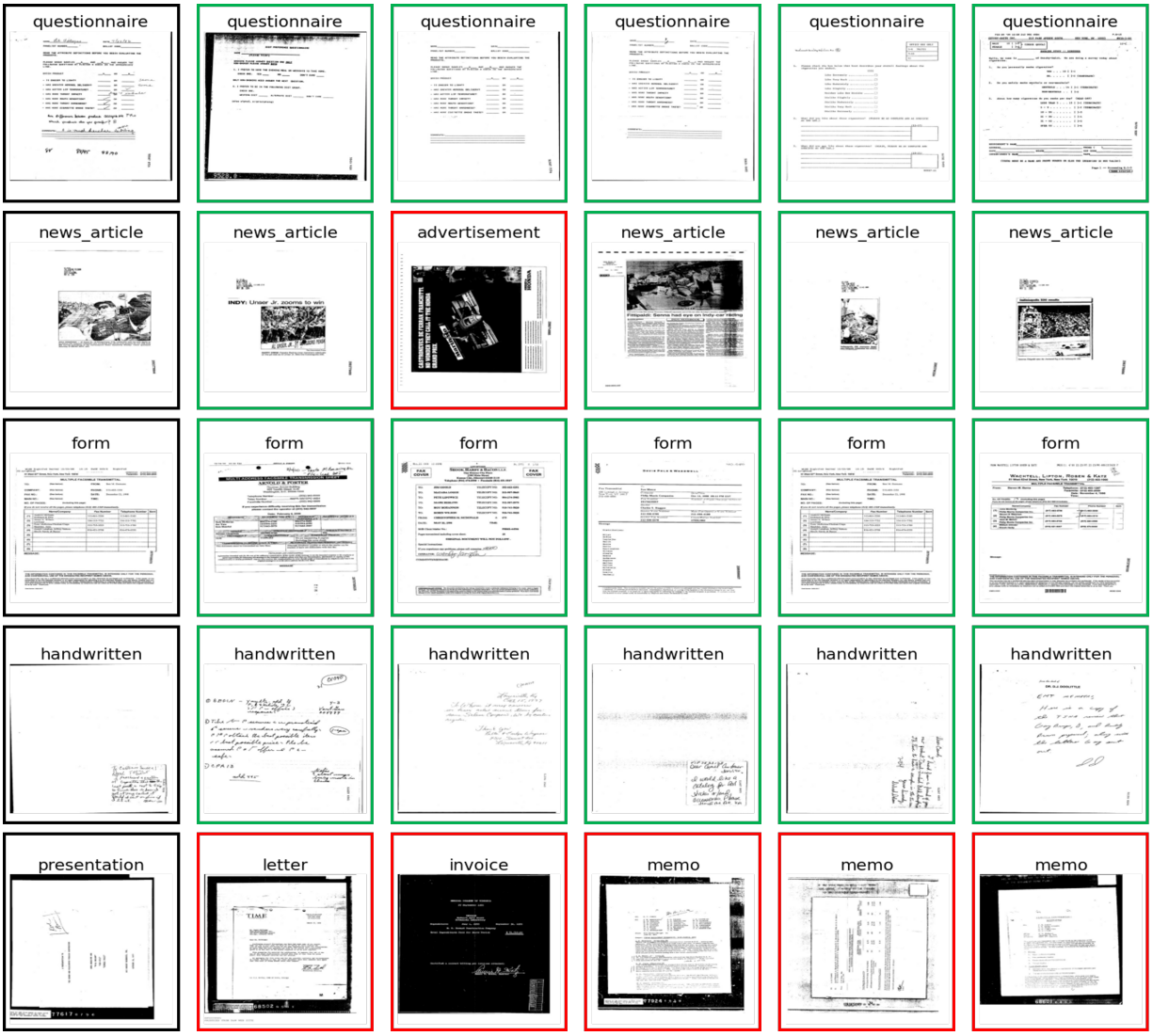



Figure 5.5: Vision to Language Representative output of the retrieval process. Randomly selected Query document images are shown in the first column, and the top-5 text sequence retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the retrieved text results for a better visualisation.

**Language → Vision.** In Figure 5.6, we retrieve the top-5 relevant document images

given a query text sequence. In the final row, we observe that the incorrectly retrieved resume document is visually similar and share the same layout information compared to the query presentation document image.

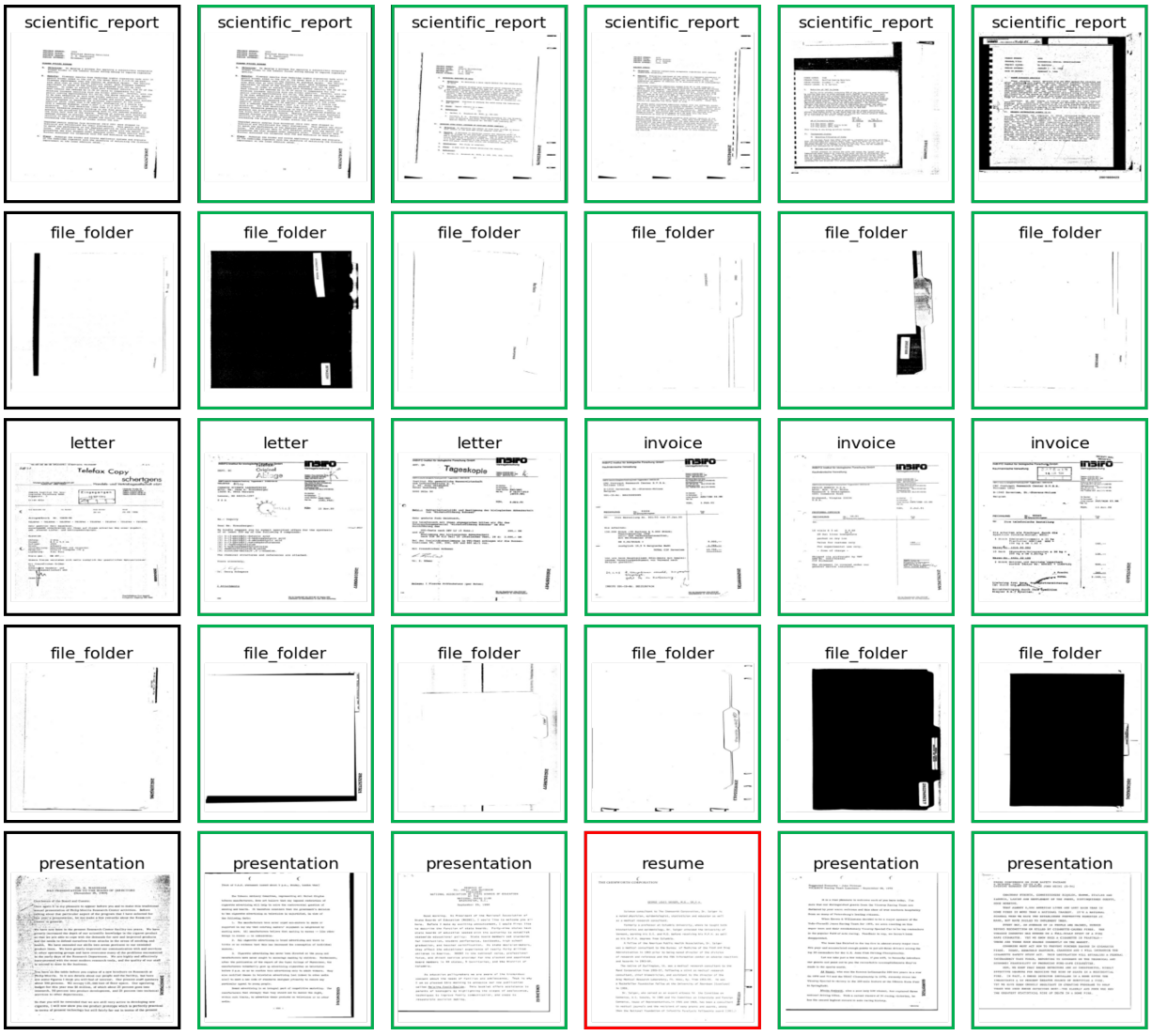



Figure 5.6: Language to Vision Representative output of the retrieval process. Randomly selected Query text sequences are shown in the first column, and the top-5 document image retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the retrieved text results for a better visualisation.

## 5.4 Discussion

In this chapter, we approached the document understanding problem by proposing a novel two-step pre-training cross-modal representation learning network, called LSRD. With the two-step pre-training approach, we first designed a novel cross-modal pre-text task, which models the intra-modality and inter-modality relations between visual and language cues using a multimodal transformer and two novel learning objectives (*i.e.* vision-language nearest-neighbor contrastive learning, and vision-language matching). In addition, we performed multimodal semantic clustering as our second pretext-task to improve the representation learning of the pre-trained embedding network. Moreover, we performed two new downstream applications in the literature of document understanding on the RVL-CDIP dataset. We evaluated the generalization ability of the learnt representations on fewer document data (*i.e.* on the few-shot classification setting), as well as its effectiveness on retrieving relevant uni-modal and cross-modal information given a query document image/text sample. Besides, we conducted the classic document classification task, which demonstrated that the gap between vision+language methods, and vision+language+layout state-of-the-art works has been narrowed. In summary, our initial goal was to build a baseline and to encourage future works in the document understanding domain to address the tasks of content-based document retrieval and few-shot document classification on the RVL-CDIP dataset.

# CHAPTER 6

# Conclusions and Future Work

The future of work consists of learning a living.

–*Marshall McLuhan*

## 6.1 Overview

In this chapter, we first summarize the contributions of this thesis to the pattern recognition and computer vision fields. In particular, its application to document understanding. Then, we highlight the main achievements and limitations of the proposed approaches. Finally, we lead the reader towards possible new research lines and natural extensions of the proposed methodologies.

## 6.2 Conclusions

In this thesis, we have introduced a study on how to classify document images using visual and textual cues incorporated within document data, by understanding and conglomerating different pattern recognition and machine learning strategies. In particular, the

huge vastness of digital document data requires a highly efficient and intuitive document understanding system. The complex layouts in which document data is often presented, poses a real challenge to vision-only-based document understanding systems. Such systems have failed to distinguish between highly correlated document categories. Though image processing research has improved significantly over the past few years, natural language research has also improved a lot in learning the semantic context of text sequences. This is where we make the assumption that jointly learning the textual content along with the visual spatial information incorporated within document data is crucial for a better understanding of multimodal document data. Although a document image is worth a thousand words, a multimodal document is worth a thousand concepts. This is where the real challenge lies to understand the nature of a document through its jointly learnt multimodal information.

## 6.3 Summary of Contributions

In this thesis, we presented the problem of multimodal document understanding. We limited our explorations to two of the well-studied modalities in the document literature, which are the vision and the language. After an introduction to the concept of multimodal document understanding and the motivation behind this thesis, we decided to make the readers familiar with the state-of-the-art of different concept and ideas for document understanding in Chapter 1. On the one hand, as an application of document understanding, we approached the document image classification problem by proposing frameworks that find a common semantic representation space for both vision and language modalities using well-known deep networks as the main backbones. Document image classification is then performed in an early feature fusion methodology (see Chapter 2). Then, we improved the semantic representation space by aligning the predictions and enabling both modalities to transfer relevant information and positive knowledge from one modality to another in a middle feature fusion manner. Document image classification is further conducted under different experimental settings (see Chapter 3). On the other hand, we also took into account the new advancements in machine learning to incorporate its strategies. We

opted to design task-agnostic and domain-agnostic pre-trained frameworks to validate the assumption that great intra-dataset generalization leads to great inter-dataset generalization. This task is performed under a pretrain-then-finetune paradigm (see Chapter 4). As well, we tackled the problem of lack of availability of human annotated document data and improved semantic representation learning by encouraging multimodal interaction within language and vision modalities in a self-supervised learning manner. We performed different ablation studies demonstrating the effectiveness of our approach on the well-established document classification task. Thus, reducing the gap with state-of-the-art works that rely on vision, language, and layout information. Also, we performed new experiments on two novel downstream tasks that we introduced as a baseline in the document understanding literature. These tasks are few-shot document classification and content-based document retrieval (see Chapter 5).

The contributions presented in this work are enumerated in four points. Moreover, even though the focus of this thesis is the development of multimodal document image classification methodologies, some of the contributions are generic algorithms applied for multimodal data in the computer vision field. Let us briefly summarize these four contributions:

- **Multimodal Deep Feature Fusion:** In Chapter 2, we proposed a two-stream deep neural network that leverages both the learned textual embeddings and visual features in an early fusion manner to classify document images. We showed that the joint learning methodology boosts the overall accuracy compared to the single-modal networks. We introduced two feature fusion methodologies to merge vision and language features in the cross-modal framework. We evaluated the performance of static and contextualized dynamic word embeddings to classify textual content of document images. As well, we reviewed the impact of training heavyweight and lightweight deep neural networks on learning relevant structural information from document images. Both the theoretical analysis and the experimental results demonstrated the superiority of our proposed joint feature learning method compared to the single-modal (*i.e.* uni-modal) modalities. This joint learning approach outperforms the state-of-the-art results with a classification accuracy of 97.05% on the large-scale

RVL-CDIP dataset, and outperforming the current state-of-the-art method by 3.91% of classification accuracy on the low-scale benchmark Tobacco-3482 dataset.

- **Multimodal Deep Mutual Learning:** In Chapter 3, we introduced a mutual learning strategy to overcome the limitations of the conventional mutual learning strategy when tested on document data. The proposed approach allowed us to learn the positive knowledge from one modality to another during the training stage, instead of the negative knowledge which we proved to weaken the learning capacity of the modality in the learning process. We presented a self-attention-based feature fusion module for a better multimodal feature extraction to perform fine-grained document image classification. Our proposed self-attention-module enhanced the overall accuracy of the ensemble network and achieved state-of-the-art classification performance compared to single-modal and multimodal methods. We performed a comprehensive ablation study on the benchmark RVL-CDIP and Tobacco-3482 datasets to analyze the effectiveness of our proposed ensemble trainable network with/without the mutual learning approach, and with/without the self-attention-based feature fusion module. We evaluated the performance and the generalization ability of the proposed ensemble network on unseen document data through inter-dataset and intra-dataset evaluation on both datasets for the single-modal and cross-modal fusion modalities. The experimental results demonstrated the effectiveness of our approach in terms of accuracy for the single-modal and cross-modal modalities. Thus, the proposed ensemble self-attention-based mutual learning model outperforms the state-of-the-art classification results based on the benchmark RVL-CDIP and Tobacco-3482 datasets.

- **Multimodal Document Representation Learning:** In Chapter 4, we designed a unified task-agnostic document pre-training framework for a better cross-modal representation learning. Our network consisted of leveraging two flexible extra levels of cross-modal interactions through cross-attention (InterMCA) and self-attention (IntraMSA) middle feature fusion-based attention modules. These modules captured high-level interactions between visual-textual cues within different document compo-

nents. We proposed a cross-modal contrastive learning objective to further explore the relations between vision and language cues. Compared to the classic single-modal contrastive learning, the proposed cross-modal contrastive learning objective allowed us to learn and align the feature representations within and across modalities. Under a fair comparison setting, our task-agnostic framework demonstrated a good generalization ability among vision and language approaches on the benchmark document datasets. It enabled us to learn robust and domain-agnostic feature representations. Thus, it achieved better results compared to the generalization experiment design conducted in Chapter 3 for the document classification task. We showed that a transformer-based architecture used in our task-agnostic pre-trained framework can achieve comparable performance when pre-trained on fewer data. The extensive experiments conducted on the public document classification datasets demonstrated the effectiveness and the generalization capacity of our model on both low-scale and large-scale datasets.

- **Improved Multimodal Semantic Document Representation Learning:** In Chapter 5, we intended to improve the semantic representation learning of our previous model introduced in chapter 3 in a self-supervised learning fashion. We introduced multimodal nearest-neighbour contrastive learning to learn self-supervised representations that go beyond single instance positives as pretext task. We showed that our network can efficiently leverage the multimodal information from unlabeled documents which benefits from modeling the interaction between language and vision modalities in the pre-training stage. Experimental evaluation showed that our network outperforms some prior works which are based on the vision-language modalities, and achieved compelling results compared to models which are based on vision, language, and layout modalities on the specific task of document classification. We addressed and explored two new downstream applications in document understanding, which are few-shot document classification and content-based document retrieval, to evaluate the effectiveness of the learnt multimodal representations to transfer to new tasks.

## 6.4 Future Research

Taking into account the lessons learnt from this work, and the improvements due to recent models that are actively being developed in the research community, we list in the following paragraphs what we identified as key topics for future research in the field of multimodal document understanding. Along the thesis, we have already stressed upon some open worth considering questions as unexplored lines that are actively being developed in the research community. Moreover, taking into account the improvement of deep learning methods, we are convinced that there is still a wide variety of research tasks for improving and advancing our work. Also, note that the new methodologies derived from the deep learning field have opened several research lines that were not covered in this dissertation. Deep learning is experiencing an evolution from the point of view of the learning strategies. The huge amount of data required for the supervision of new models causes a huge bottleneck dealing with new problems. Therefore, self-supervised learning strategies are gaining popularity among the machine learning community, and more specifically, among the document understanding community in the last three years. Taking into account the outcomes of this dissertation, there are many extensions that can be made. We list in the following paragraphs key topics for future research in the field of multimodal document understanding.

**Multimodal Fusion and Reasoning.** In this thesis, we explored several multimodal fusion techniques to leverage information from vision and language cues. However, in the multimodal machine learning literature, there have been several studies on designing models capable of reasoning to explore the synergy between visual and textual features [102, 197] in a sequential manner. Also, significant advances have been made by the use of Graph convolutional networks (GCN) [82] which are gaining importance in many multimodal tasks such as image captioning [105], image-sentence retrieval [107], and visual question answering (VQA) [124]. GCN are able to model relationships between nodes in a given graph and to explore semantic correlation between visual and textual features [116, 153]. Therefore, as a first future line of research, we aim to combine textual features with salient image regions in document images to exploit the complementary information

carried by the two sources. Specifically, we will employ a Graph Convolutional Network (GCN) to perform multimodal reasoning and obtain relationship-enhanced features by learning a common semantic space between salient image regions and text sequences in document images.

**Multimodal Document Understanding with GNNs.** Graph reasoning has been recently applied to document understanding tasks such as key-information extraction [27], document layout analysis [146], table structure recognition [108], table extraction [54], visual question answering [106], and synthetic document generation [24], etc. In the future work, we intend to use the power of graphs in representing: (1) the spatial structure of document images with usage of the positional information of object categories like tables, titles, figures; (2) the semantic conceptual connections between the different object categories in a document (*e.g.* recognizing the semantic text entities and their relationships from documents). Therefore, we will study the impact of leveraging graph representations as a third modality in our proposed task-agnostic pretrained framework in Chapter 5, on enhancing the quality of document representation. However, as there exists no positional information in the RVL-CDIP and Tobacco-3482 datasets, we will explore the heavy-scale document datasets (*i.e.* Industry document dataset (IDL) which consists of 26M documents with OCR Annotations). Such document understanding system will be able to generalize better on unseen data, and thus, can be transferred to other domain-specific multimodal data. Hence, deriving an off-the-shelf document analysis solution, to be performed on various document downstream tasks that we have not explored before in this thesis, which are: Document (DoCVQA), form and receipt understanding,sequence labeling, and also document layout detection.

**Synergistic Learning between Multiple Modalities/Domains.** Given the heterogeneity and variability of complex layouts and graphical entities incorporated within document data, it poses a great challenge to deep CNNs and transformers to distinguish between highly correlated documents. Despite huge vision-language model pre-training methods achieving superior performance on most multimodal document understanding tasks, large-scale document pre-training comes with a high computational cost both in terms of memory and training time. Therefore, as synergistic learning is one of the future

research lines we would like to explore, we principally aim to view each document category as a unique modality/domain. We want the model to learn more specific information about each category of document data in an incremental learning manner. Then, mutual learning can be introduced to transfer the information learnt within each modality/domain. We believe this approach will help the model to learn more relevant information that is hard to be learnt in the case where the model is given all document data at once.

# Appendices

# APPENDIX A

# Appendix

## Additional Results: Evaluation On Uni-Modal Content-based Document Retrieval

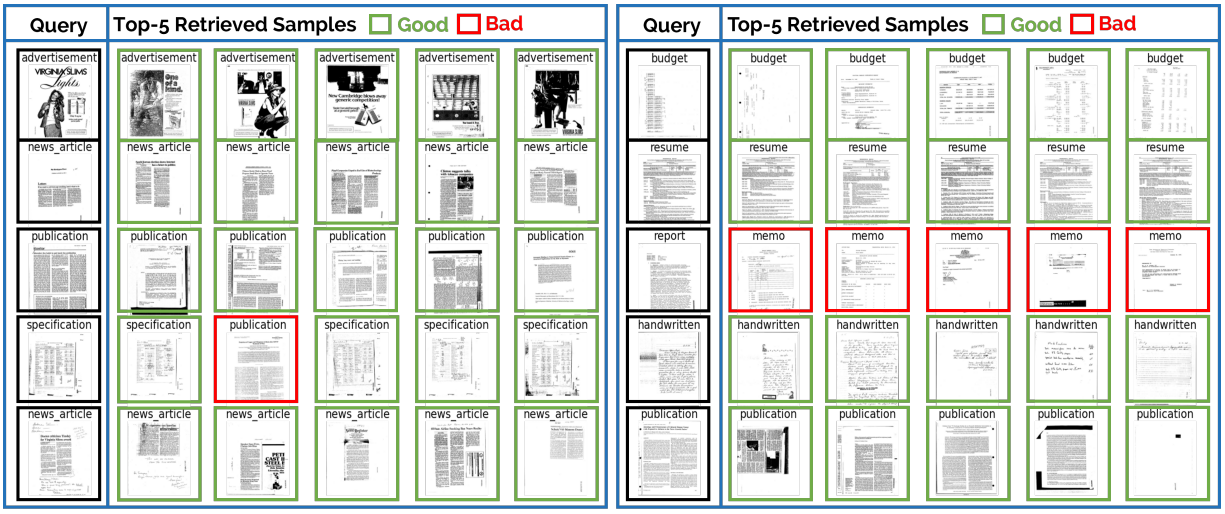



Figure A.1: Vision to Vision Representative output of the retrieval process. Randomly selected Query document images are shown in the first column, and the top-5 document image retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border.

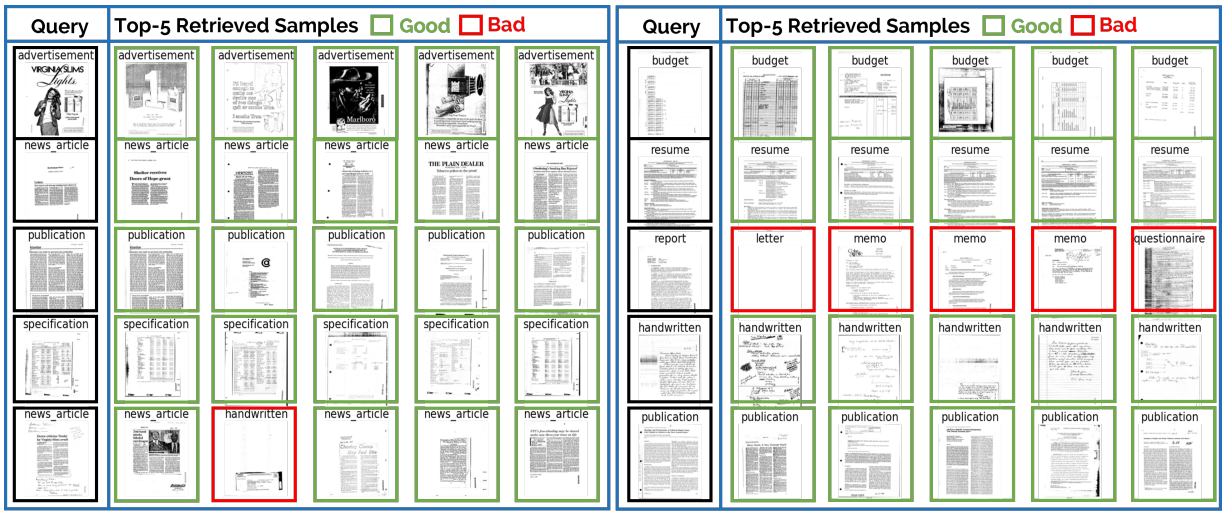



Figure A.2: Language to Language Representative output of the retrieval process. Randomly selected Text sequences are used as query in the first column, and the top-5 text sequence retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the queries and retrieved results for a better visualisation.

# APPENDIX B

# Appendix

## Additional Results: Evaluation On Cross-Modal Content-based Document Retrieval

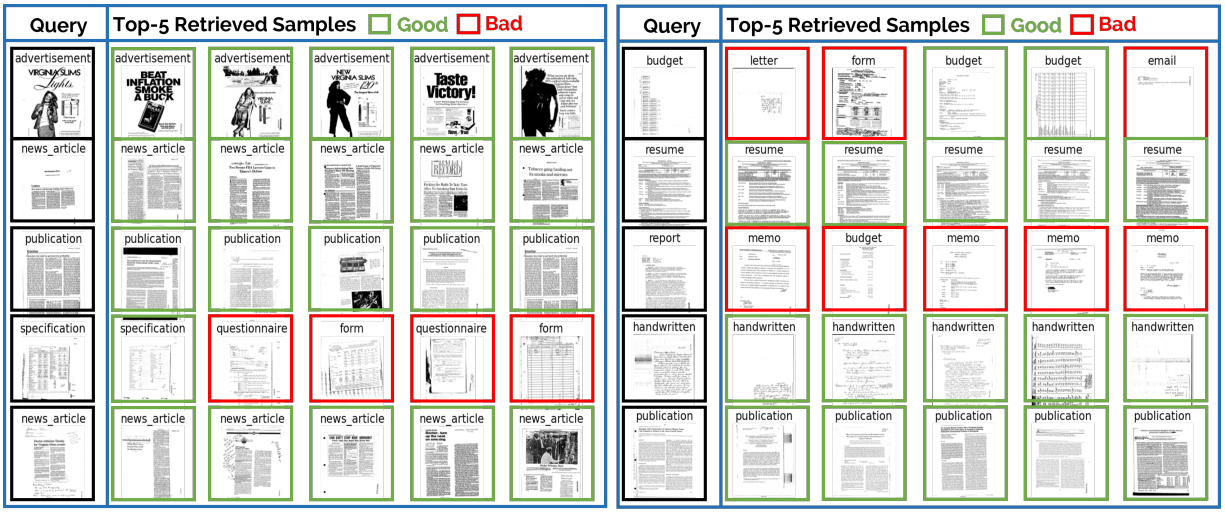



Figure B.1: Vision to Language Representative output of the retrieval process. Randomly selected Query document images are shown in the first column, and the top-5 text sequence retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. The corresponding document images of the retrieved text results are shown.

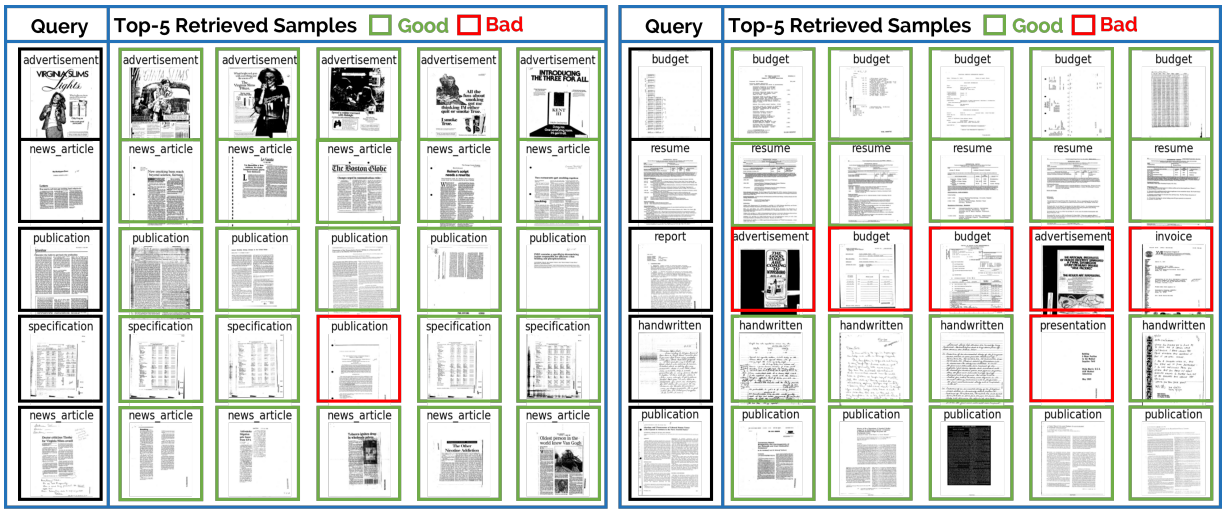



Figure B.2: Language to Vision Representative output of the retrieval process. Randomly selected Query text sequences are shown in the first column, and the top-5 document image retrievals are shown in the following columns in order. Retrievals from the same class are shown with a green border; retrievals from a different class are shown with a red border. We show the corresponding document images of the retrieved text results for a better visualisation.

# Bibliography


[1] Sherif Abuelwafa, Marco Pedersoli, and Mohamed Cheriet. "Unsupervised exemplar-based learning for improved document image classification". In: *IEEE Access* 7 (2019), pp. 133738–133748.

[2] M. Afzal, Joan Pastor-Pellicer, F. Shafait, T. Breuel, Andreas Dengel, and Marcus Liwicki. "Document Image Binarization using LSTM: A Sequence Learning Approach". In: *HIP '15*. 2015.

[3] Muhammad Zeshan Afzal, Samuele Capobianco, Muhammad Imran Malik, Simone Marinai, Thomas M Breuel, Andreas Dengel, and Marcus Liwicki. "Deepdocclassifier: Document classification with deep convolutional neural network". In: *2015 13th international conference on document analysis and recognition (ICDAR)*. IEEE. 2015, pp. 1111–1115.

[4] Muhammad Zeshan Afzal, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. "Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 883–888.

[5] Madhav Agarwal, Ajoy Mondal, and CV Jawahar. "Cdec-net: Composite deformable cascade network for table detection in document images". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 9491–9498.

[6] Mohamed Aly, Peter Welinder, Mario Munich, and Pietro Perona. "Automatic discovery of image families: Global vs. local features". In: *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2009, pp. 777–780.

[7] Shun-ichi Amari. "Backpropagation and stochastic gradient descent method". In: *Neurocomputing* 5.4-5 (1993), pp. 185–196.

[8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6077–6086.

[9] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. "Docformer: End-to-end transformer for document understanding". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 993–1003.

[10] Enrico Appiani, Francesca Cesarini, Anna Maria Colla, Michelangelo Diligenti, Marco Gori, Simone Marinai, and Giovanni Soda. "Automatic document classification and indexing in high-volume applications". In: *International Journal on Document Analysis and Recognition* 4.2 (2001), pp. 69–83.

[11] Muhammad Nabeel Asim, Muhammad Usman Ghani Khan, Muhammad Imran Malik, Khizar Razzaque, Andreas Dengel, and Sheraz Ahmed. "Two stream deep network for document image classification". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 1410–1416.

[12] Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. "Multimodal deep networks for text and image-based document classification". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, pp. 427–443.

[13] Olivier Augereau, N. Journet, Anne Vialard, and Jean-Philippe Domenger. "Improving Classification of an Industrial Document Image Database by Combining Visual and Textual Features". In: *2014 11th IAPR International Workshop on Document Analysis Systems* (2014), pp. 314–318.

[14] Jimmy Ba and Rich Caruana. "Do deep nets really need to be deep?" In: *Advances in neural information processing systems* 27 (2014).

[15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[16] Souhail Bakkali, Zuheng Ming, Mickaël Coustaty, and Marçal Rusiñol. "Cross-modal deep networks for document image classification". In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 2556–2560.

[17] Souhail Bakkali, Zuheng Ming, Mickaël Coustaty, and Marçal Rusiñol. "Visual and textual deep feature fusion for document image classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 562–563.

[18] Souhail Bakkali, Zuheng Ming, Mickaël Coustaty, and Marçal Rusiñol. "EAML: ensemble self-attention-based mutual learning network for document image classification". In: *International Journal on Document Analysis and Recognition (IJDAR)* 24.3 (2021), pp. 251–268.

[19] Souhail Bakkali, Zuheng Ming, Mickael Coustaty, Marçal Rusiñol, and Oriol Ramos Terrades. "VLCDoC: Vision-Language Contrastive Pre-Training Model for Cross-Modal Document Classification". In: *arXiv preprint arXiv:2205.12029* (2022).

[20] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy". In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018), pp. 423–443.

[21] Hangbo Bao, Li Dong, and Furu Wei. "Beit: Bert pre-training of image transformers". In: *arXiv preprint arXiv:2106.08254* (2021).

[22] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features". In: *European conference on computer vision*. Springer. 2006, pp. 404–417.

[23] Serge Belongie, Jitendra Malik, and Jan Puzicha. "Shape matching and object recognition using shape contexts". In: *IEEE transactions on pattern analysis and machine intelligence* 24.4 (2002), pp. 509–522.

[24] Sanket Biswas, Pau Riba, Josep Lladós, and Umapada Pal. "Graph-Based Deep Generative Modelling for Document Layout Generation". In: *International Conference on Document Analysis and Recognition*. Springer. 2021, pp. 525–537.

[25] Piotr Bojanowski and Armand Joulin. "Unsupervised learning by predicting noise". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 517–526.

[26] Yungcheol Byun and Yillbyung Lee. "Form classification using DP matching". In: *Proceedings of the 2000 ACM symposium on Applied computing-Volume 1*. 2000, pp. 1–4.

[27] Manuel Carbonell, Pau Riba, Mauricio Villegas, Alicia Fornés, and Josep Lladós. "Named entity recognition and relation extraction with graph neural networks in semi structured documents". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 9622–9627.

[28] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep clustering for unsupervised learning of visual features". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149.

[29] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. "Abc-cnn: An attention based convolutional neural network for visual question answering". In: *arXiv preprint arXiv:1511.05960* (2015).

[30] Nawei Chen and Dorothea Blostein. "A survey of document image classification: problem statement, classifier architecture and performance evaluation". In: *International Journal of Document Analysis and Recognition (IJDAR)* 10.1 (2007), pp. 1–16.

[31] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[32] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. "A closer look at few-shot classification". In: *arXiv preprint arXiv:1904.04232* (2019).

[33] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[34] Gabriela Csurka. "Document image classification, with a specific view on applications of patent images". In: *Current Challenges in Patent Information Retrieval*. Springer, 2017, pp. 325–350.

[35] Gabriela Csurka, Diane Larlus, Albert Gordo, and Jon Almazan. "What is the right way to represent document images?" In: *arXiv preprint arXiv:1603.01076* (2016).

[36] Arindam Das, Saikat Roy, Ujjwal Bhattacharya, and Swapan K Parui. "Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks". In: *2018 24th international conference on pattern recognition (ICPR)*. IEEE. 2018, pp. 3180–3185.

[37] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. "Align2ground: Weakly supervised phrase grounding guided by image-caption alignment". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2601–2610.

[38] T. Dauphinee, N. Patel, and Mohammad Mehdi Rashidi. "Modular Multimodal Architecture for Document Classification". In: *ArXiv* abs/1912.04376 (2019).

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[40] Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. "Bag-of-visual-words models for adult image classification and filtering". In: *2008 19th International Conference on Pattern Recognition*. IEEE. 2008, pp. 1–4.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[42] Terrance DeVries and Graham W Taylor. "Improved regularization of convolutional neural networks with cutout". In: *arXiv preprint arXiv:1708.04552* (2017).

[43] Carl Doersch, Abhinav Gupta, and Alexei A Efros. "Unsupervised visual representation learning by context prediction". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1422–1430.

[44] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. "Adversarial feature learning". In: *arXiv preprint arXiv:1605.09782* (2016).

[45] Jeff Donahue and Karen Simonyan. "Large scale adversarial representation learning". In: *Advances in neural information processing systems* 32 (2019).

[46] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[47] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 9588–9597.

[48] Javier Ferrando, Juan Luis Domínguez, Jordi Torres, Raúl García, David García, Daniel Garrido, Jordi Cortada, and Mateo Valero. "Improving accuracy and speeding up document image classification through parallel systems". In: *International Conference on Computational Science*. Springer. 2020, pp. 387–400.

[49] Chelsea Finn, Pieter Abbeel, and Sergey Levine. “Model-agnostic meta-learning for fast adaptation of deep networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 1126–1135.

[50] George Forman et al. “An extensive empirical study of feature selection metrics for text classification.” In: *J. Mach. Learn. Res.* 3.Mar (2003), pp. 1289–1305.

[51] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. “Multimodal compact bilinear pooling for visual question answering and visual grounding”. In: *arXiv preprint arXiv:1606.01847* (2016).

[52] I. Gallo, Alessandro Calefati, S. Nawaz, and Muhammad Kamran Janjua. “Image and Encoded Text Fusion for Multi-Modal Classification”. In: *2018 Digital Image Computing: Techniques and Applications (DICTA)* (2018), pp. 1–7.

[53] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. “A survey on deep learning for multimodal data fusion”. In: *Neural Computation* 32.5 (2020), pp. 829–864.

[54] Andrea Gemelli, Emanuele Vivoli, and Simone Marinai. “Graph neural networks and representation embedding for table extraction in PDF documents”. In: *arXiv preprint arXiv:2208.11203* (2022).

[55] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. “Unsupervised representation learning by predicting image rotations”. In: *arXiv preprint arXiv:1803.07728* (2018).

[56] Yoav Goldberg and Omer Levy. “word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method”. In: *arXiv preprint arXiv:1402.3722* (2014).

[57] Albert Gordo, Florent Perronnin, and Ernest Valveny. “Large-scale document image retrieval and classification with runlength histograms and binary embeddings”. In: *Pattern Recognition* 46.7 (2013), pp. 1898–1905.

[58] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. “LSTM: A search space odyssey”. In: *IEEE transactions on neural networks and learning systems* 28.10 (2016), pp. 2222–2232.

[59] Jiuxiang Gu, Jason Kuen, Shafiq Joty, Jianfei Cai, Vlad Morariu, Handong Zhao, and Tong Sun. "Self-Supervised Relationship Probing". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1841–1853.

[60] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, and Tong Sun. "Unified Pretraining Framework for Document Understanding". In: *arXiv preprint arXiv:2204.10939* (2022).

[61] Adam W. Harley, A. Ufkes, and K. Derpanis. "Evaluation of deep convolutional nets for document image classification and retrieval". In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (2015), pp. 991–995.

[62] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[64] Olivier Henaff. "Data-efficient image recognition with contrastive predictive coding". In: *International conference on machine learning*. PMLR. 2020, pp. 4182–4192.

[65] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531* (2015).

[66] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. "A comprehensive survey of deep learning for image captioning". In: *ACM Computing Surveys (CsUR)* 51.6 (2019), pp. 1–36.

[67] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[68] Rui Hu and John Collomosse. "A performance evaluation of gradient field hog descriptor for sketch based image retrieval". In: *Computer Vision and Image Understanding* 117.7 (2013), pp. 790–806.

[69] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[70] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: a new learning scheme of feedforward neural networks". In: *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*. Vol. 2. Ieee. 2004, pp. 985–990.

[71] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications". In: *Neurocomputing* 70.1-3 (2006), pp. 489–501.

[72] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking". In: *arXiv preprint arXiv:2204.08387* (2022).

[73] Noman Islam, Zeeshan Islam, and Nazia Noor. "A survey on optical character recognition system". In: *arXiv preprint arXiv:1710.05703* (2017).

[74] Anil K Jain. "Data clustering: 50 years beyond K-means". In: *Pattern recognition letters* 31.8 (2010), pp. 651–666.

[75] Simon Jenni and Paolo Favaro. "Self-supervised feature learning by learning to spot artifacts". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2733–2742.

[76] Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann. "Convolutional neural networks for document image classification". In: *2014 22nd International Conference on Pattern Recognition*. IEEE. 2014, pp. 3168–3172.

[77] Abdul Amir Abdullah Karim and Rafal Ali Sameer. "Image classification using bag of visual words (bovw)". In: *Al-Nahrain Journal of Science* 21.4 (2018), pp. 76–82.

[78] Yan Ke and Rahul Sukthankar. "PCA-SIFT: A more distinctive representation for local image descriptors". In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Vol. 2. IEEE. 2004, pp. II–II.

[79] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. "Bilinear attention networks". In: *Advances in neural information processing systems* 31 (2018).

[80] Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee. "Deep CNN-based blind image quality predictor". In: *IEEE transactions on neural networks and learning systems* 30.1 (2018), pp. 11–24.

[81] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[82] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).

[83] Florian Kleber, Markus Diem, and Robert Sablatnig. "Form classification and retrieval using bag of words with shape features of line structures". In: *Document Recognition and Retrieval XXI*. Vol. 9021. SPIE. 2014, pp. 61–69.

[84] Takumi Kobayashi. "BFO meets HOG: feature extraction based on histograms of oriented pdf gradients for image classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 747–754.

[85] Andreas Kölsch, Muhammad Zeshan Afzal, Markus Ebbecke, and Marcus Liwicki. "Real-time document image classification using deep CNN and extreme learning machines". In: *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 1318–1323.

[86] Praveen Krishnan and CV Jawahar. "Matching handwritten document images". In: *European Conference on Computer Vision*. Springer. 2016, pp. 766–782.

[87] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[88] Jayant Kumar and David Doermann. "Unsupervised classification of structurally similar document images". In: *2013 12th International Conference on Document Analysis and Recognition*. IEEE. 2013, pp. 1225–1229.

[89] Jayant Kumar, Peng Ye, and David Doermann. "Learning document structure for retrieval and classification". In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE. 2012, pp. 1558–1561.

[90] Jayant Kumar, Peng Ye, and David S. Doermann. "Structural similarity for document image classification and retrieval". In: *Pattern Recognit. Lett.* 43 (2014), pp. 119–126.

[91] Dana Lahat, Tülay Adali, and Christian Jutten. "Multimodal data fusion: an overview of methods, challenges, and prospects". In: *Proceedings of the IEEE* 103.9 (2015), pp. 1449–1477.

[92] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. "Recurrent convolutional neural networks for text classification". In: *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[93] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. "Human-level concept learning through probabilistic program induction". In: *Science* 350.6266 (2015), pp. 1332–1338.

[94] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. "Colorization as a proxy task for visual understanding". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6874–6883.

[95] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". In: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. Vol. 2. IEEE. 2006, pp. 2169–2178.

[96] Viet Phuong Le, Muriel Visani, Cao De Tran, and Jean-Marc Ogier. "Logo spotting for document categorization". In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE. 2012, pp. 3484–3487.

[97] Viet Phuong Le, Muriel Visani, Cao De Tran, and Jean-Marc Ogier. "Improving logo spotting and matching for document categorization by a post-filter based on homography". In: *2013 12th International Conference on Document Analysis and Recognition*. IEEE. 2013, pp. 270–274.

[98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[99] Yann LeCun, Larry Jackel, Leon Bottou, A Brunot, Corinna Cortes, John Denker, Harris Drucker, Isabelle Guyon, UA Muller, Eduard Sackinger, et al. "Comparison of learning algorithms for handwritten digit recognition". In: *International conference on artificial neural networks*. Vol. 60. 1. Perth, Australia. 1995, pp. 53–60.

[100] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. "Stacked cross attention for image-text matching". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 201–216.

[101] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. "DiT: Self-supervised Pre-training for Document Image Transformer". In: *arXiv preprint arXiv:2203.02378* (2022).

[102] Kunpeng Li, Yulun Zhang, K. Li, Yuanyuan Li, and Yun Fu. "Visual Semantic Reasoning for Image-Text Matching". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 4653–4661.

[103] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. "Visualbert: A simple and performant baseline for vision and language". In: *arXiv preprint arXiv:1908.03557* (2019).

[104] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. "Selfdoc: Self-supervised document representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5652–5660.

[105] Xiangyang Li and Shuqiang Jiang. "Know more say less: Image captioning based on scene graphs". In: *IEEE Transactions on Multimedia* 21.8 (2019), pp. 2117–2130.

[106] Yaoyuan Liang, Xin Wang, Xuguang Duan, and Wenwu Zhu. "Multi-modal Contextual Graph Neural Network for Text Visual Question Answering". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 3491–3498.

[107] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. "Graph structured network for image-text matching". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10921–10930.

[108] Hao Liu, Xin Li, Bing Liu, Deqiang Jiang, Yinsong Liu, and Bo Ren. "Neural Collaborative Graph Machines for Table Structure Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4533–4542.

[109] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. "Graph convolution for multimodal information extraction from visually rich documents". In: *arXiv preprint arXiv:1903.11279* (2019).

[110] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[111] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. "Learning a recurrent residual fusion network for multimodal matching". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4107–4116.

[112] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[113] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[114] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks". In: *Advances in neural information processing systems* 32 (2019).

[115] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Hierarchical question-image co-attention for visual question answering". In: *Advances in neural information processing systems* 29 (2016).

[116] Andres Mafla, Sounak Dey, Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. "Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 4023–4033.

[117] Krystian Mikolajczyk and Cordelia Schmid. "A performance evaluation of local descriptors". In: *IEEE transactions on pattern analysis and machine intelligence* 27.10 (2005), pp. 1615–1630.

[118] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[119] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. "Advances in Pre-Training Distributed Word Representations". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[120] Marcin Michał Mirończuk and Jarosław Protasiewicz. "A recent overview of the state-of-the-art elements of text classification". In: *Expert Systems with Applications* 106 (2018), pp. 36–54.

[121] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6707–6717.

[122] Kosuke Mizuno, Yosuke Terachi, Kenta Takagi, Shintaro Izumi, Hiroshi Kawaguchi, and Masahiko Yoshimoto. "Architectural study of HOG feature extraction proces-

sor for real-time object detection". In: *2012 IEEE Workshop on Signal Processing Systems*. IEEE. 2012, pp. 197–202.

[123] T Nathan Mundhenk, Daniel Ho, and Barry Y Chen. "Improvements to context based self-supervised learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9339–9348.

[124] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. "Out of the box: Reasoning with graph convolution nets for factual visual question answering". In: *Advances in neural information processing systems* 31 (2018).

[125] Duy-Kien Nguyen and Takayuki Okatani. "Improved Fusion of Visual and Language Representations by Dense Symmetric Co-attention for Visual Question Answering". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 6087–6096.

[126] Lucia Noce, Ignazio Gallo, Alessandro Zamberletti, and Alessandro Calefati. "Embedded textual content for document image classification with convolutional neural networks". In: *Proceedings of the 2016 ACM Symposium on Document Engineering*. 2016, pp. 165–173.

[127] Mehdi Noroozi and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *European conference on computer vision*. Springer. 2016, pp. 69–84.

[128] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. "Representation learning by learning to count". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5898–5906.

[129] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. "Boosting self-supervised learning via knowledge transfer". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9359–9367.

[130] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[131] Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. "Actor-mimic: Deep multitask and transfer reinforcement learning". In: *arXiv preprint arXiv:1511.06342* (2015).

[132] Joan Pastor-Pellicer, M. Afzal, Marcus Liwicki, and M. J. Bleda. "Complete System for Text Line Extraction Using Convolutional Neural Networks and Watershed Transform". In: *2016 12th IAPR Workshop on Document Analysis Systems (DAS)* (2016), pp. 30–35.

[133] Joan Pastor-Pellicer, Salvador España Boquera, Francisco Zamora-Martínez, Muhammad Zeshan Afzal, and María José Castro Bleda. "Insights on the Use of Convolutional Neural Networks for Document Image Binarization". In: *IWANN*. 2015.

[134] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. "Context encoders: Feature learning by inpainting". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544.

[135] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[136] Florent Perronnin and Christopher Dance. "Fisher kernels on visual vocabularies for image categorization". In: *2007 IEEE conference on computer vision and pattern recognition*. IEEE. 2007, pp. 1–8.

[137] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations". In: *NAACL*. 2018.

[138] 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". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 6892–6899.

[139] Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. "Going full-tilt boogie on document understanding with text-image-layout transformer". In: *International Conference on Document Analysis and Recognition*. Springer. 2021, pp. 732–747.

[140] Jianjun Qian, Weilan Wang, and Daohui Wang. "A Novel Approach for Online Handwriting Recognition of Tibetan Characters". In: 2010.

[141] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. "Few-shot image recognition by predicting parameters from activations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7229–7238.

[142] Sachin Raja, Ajoy Mondal, and CV Jawahar. "Visual Understanding of Complex Table Structures from Document Images". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 2299–2308.

[143] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. "Stand-Alone Self-Attention in Vision Models". In: (2019). arXiv: 1906.05909 [cs.CV].

[144] Sachin Ravi and Hugo Larochelle. "Optimization as a model for few-shot learning". In: (2016).

[145] Zhongzheng Ren and Yong Jae Lee. "Cross-domain self-supervised multi-task feature learning using synthetic imagery". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 762–771.

[146] Pau Riba, Anjan Dutta, Lutz Goldmann, Alicia Fornés, Oriol Ramos, and Josep Lladós. "Table detection in invoice documents by graph neural networks". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 122–127.

[147] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. "Fitnets: Hints for thin deep nets". In: *arXiv preprint arXiv:1412.6550* (2014).

[148] Saikat Roy, Arindam Das, and Ujjwal Bhattacharya. "Generalized stacking of layerwise-trained deep convolutional neural networks for document image classification". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 1273–1278.

[149] Marçal Rusinol and Josep Llados. "Logo spotting by a bag-of-words approach for document categorization". In: *2009 10th international conference on document analysis and recognition*. IEEE. 2009, pp. 111–115.

[150] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[151] Prateek Sarkar. "Image classification: Classifying distributions of visual features". In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 2. IEEE. 2006, pp. 472–475.

[152] Ritesh Sarkhel and Arnab Nandi. "Deterministic routing between layout abstractions for multi-scale classification of visually rich documents". In: *28th International Joint Conference on Artificial Intelligence (IJCAI), 2019*. 2019.

[153] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. "Modeling relational data with graph convolutional networks". In: *European semantic web conference*. Springer. 2018, pp. 593–607.

[154] Paul Scovanner, Saad Ali, and Mubarak Shah. "A 3-dimensional sift descriptor and its application to action recognition". In: *Proceedings of the 15th ACM international conference on Multimedia*. 2007, pp. 357–360.

[155] Mathias Seuret, M. Alberti, Marcus Liwicki, and R. Ingold. "PCA-Initialized Deep Neural Networks Applied to Document Image Analysis". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 01 (2017), pp. 877–882.

[156] Christian K. Shin and David S. Doermann. “Document Image Retrieval Based on Layout Structural Similarity”. In: *IPCV*. 2006.

[157] Sebastián Sierra and Fabio A. González. “Combining Textual and Visual Representations for Multimodal Author Profiling: Notebook for PAN at CLEF 2018”. In: *CLEF*. 2018.

[158] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[159] Josef Sivic and Andrew Zisserman. “Video Google: A text retrieval approach to object matching in videos”. In: *Computer Vision, IEEE International Conference on*. Vol. 3. IEEE Computer Society. 2003, pp. 1470–1470.

[160] Ray Smith. “An overview of the Tesseract OCR engine”. In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. IEEE. 2007, pp. 629–633.

[161] Jake Snell, Kevin Swersky, and Richard Zemel. “Prototypical networks for few-shot learning”. In: *Advances in neural information processing systems* 30 (2017).

[162] Kihyuk Sohn. “Improved deep metric learning with multi-class n-pair loss objective”. In: *Advances in neural information processing systems* 29 (2016).

[163] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. “Learning to compare: Relation network for few-shot learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1199–1208.

[164] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. “Inception-v4, inception-resnet and the impact of residual connections on learning”. In: *Thirty-first AAAI conference on artificial intelligence*. 2017.

[165] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[166] Chris Tensmeyer and Tony Martinez. "Analysis of convolutional neural networks for document image classification". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 388–393.

[167] Chunwei Tian, Yong Xu, and Wangmeng Zuo. "Image denoising using deep CNN with batch renormalization". In: *Neural Networks* 121 (2020), pp. 461–473.

[168] Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive multiview coding". In: *European conference on computer vision*. Springer. 2020, pp. 776–794.

[169] Oliver Tüselmann, Friedrich Müller, Fabian Wolf, and Gernot A Fink. "Recognition-free Question Answering on Handwritten Document Collections". In: *arXiv preprint arXiv:2202.06080* (2022).

[170] Adnan Ul-Hasan, M. Afzal, F. Shafait, Marcus Liwicki, and T. Breuel. "A sequence learning approach for multiple script identification". In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (2015), pp. 1046–1050.

[171] Sergey Usilin, Dmitry Nikolaev, Vassili Postnikov, and Gerald Schaefer. "Visual appearance based document image classification". In: *2010 IEEE International Conference on Image Processing*. IEEE. 2010, pp. 2133–2136.

[172] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[173] Yashaswi Verma, Abhishek Jha, and CV Jawahar. "Cross-specificity: modelling data semantics for cross-modal matching and retrieval". In: *International journal of multimedia information retrieval* 7.2 (2018), pp. 139–146.

[174] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. "Matching networks for one shot learning". In: *Advances in neural information processing systems* 29 (2016).

[175] Di Wang, Quan Wang, Lihuo He, Xinbo Gao, and Yumin Tian. "Joint and individual matrix factorization hashing for large-scale cross-modal retrieval". In: *Pattern Recognition* 107 (2020), p. 107479.

[176] Jiapeng Wang, Lianwen Jin, and Kai Ding. "Lilt: A simple yet effective language-independent layout transformer for structured document understanding". In: *arXiv preprint arXiv:2202.13669* (2022).

[177] K Wang, Q Yin, W Wang, S Wu, and L Wang. "A comprehensive survey on cross-modal retrieval (2016)". In: *arXiv preprint arXiv:1607.06215* ().

[178] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. "Learning two-branch neural networks for image-text matching tasks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2018), pp. 394–407.

[179] X. Wang, Ross B. Girshick, A. Gupta, and Kaiming He. "Non-local Neural Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7794–7803.

[180] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. "Position focused attention network for image-text matching". In: *arXiv preprint arXiv:1907.09748* (2019).

[181] Fei Wu, Xiao-Yuan Jing, Zhiyong Wu, Yimu Ji, Xiwei Dong, Xiaokai Luo, Qinghua Huang, and Ruchuan Wang. "Modality-specific and shared generative adversarial network for cross-modal retrieval". In: *Pattern Recognition* 104 (2020), p. 107335.

[182] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. "Unsupervised feature learning via non-parametric instance discrimination". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3733–3742.

[183] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1492–1500.

[184] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. "LayoutLMv2: Multi-modal pre-training for visually-rich document understanding". In: *arXiv preprint arXiv:2012.14740* (2020).

[185] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. "Layoutlm: Pre-training of text and layout for document image understanding". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1192–1200.

[186] Shiyang Yan, Yuan Xie, F. Wu, J. Smith, Wenjin Lu, and B. Zhang. "Image captioning via hierarchical attention mechanism and policy gradient optimization". In: *Signal Process.* 167 (2020).

[187] Fan Yang, Lianwen Jin, Weixin Yang, Ziyong Feng, and Shuye Zhang. "Handwritten/printed receipt classification using attention-based convolutional neural network". In: *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE. 2016, pp. 384–389.

[188] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. "Exploring deep multimodal fusion of text and photo for hate speech classification". In: *Proceedings of the third workshop on abusive language online*. 2019, pp. 11–18.

[189] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. "Learning to extract semantic structure from documents using multimodal fully convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5315–5324.

[190] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* 32 (2019).

[191] Zichao Yang, X. He, Jianfeng Gao, L. Deng, and Alex Smola. "Stacked Attention Networks for Image Question Answering". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 21–29.

[192] Zhou Yu, Yuhao Cui, Jun Yu, Dacheng Tao, and Qi Tian. "Multimodal unified attention networks for vision-and-language interactions". In: *arXiv preprint arXiv:1908.04107* (2019).

[193] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1821–1830.

[194] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering". In: *IEEE transactions on neural networks and learning systems* 29.12 (2018), pp. 5947–5959.

[195] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. "Multimodal contrastive training for visual representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 6995–7004.

[196] Tom Zahavy, Alessandro Magnani, Abhinandan Krishnan, and Shie Mannor. "Is a picture worth a thousand words? A Deep Multi-Modal Fusion Architecture for Product Classification in e-commerce". In: *arXiv preprint arXiv:1611.09534* (2016).

[197] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. "From recognition to cognition: Visual commonsense reasoning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 6720–6731.

[198] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. "Cross-modal contrastive learning for text-to-image generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 833–842.

[199] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. "Learning deep CNN denoiser prior for image restoration". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3929–3938.

[200] Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. "TRIE: end-to-end text reading and information extraction for document understanding". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 1413–1422.

[201] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization". In: *European conference on computer vision*. Springer. 2016, pp. 649–666.

[202] Xiang Zhang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification". In: *Advances in neural information processing systems* 28 (2015).

[203] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. "Deep mutual learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4320–4328.

[204] Hengshuang Zhao, Jiaya Jia, and V. Koltun. "Exploring Self-Attention for Image Recognition". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 10073–10082.

[205] Rui Zhao and Kezhi Mao. "Fuzzy bag-of-words model for document representation". In: *IEEE transactions on fuzzy systems* 26.2 (2017), pp. 794–804.

[206] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. "Deep supervised cross-modal retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10394–10403.

[207] B. Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and R. Fergus. "Simple Baseline for Visual Question Answering". In: *ArXiv* abs/1512.02167 (2015).

[208] Huiyu Zhou, Yuan Yuan, and Chunmei Shi. "Object tracking using SIFT features and mean shift". In: *Computer vision and image understanding* 113.3 (2009), pp. 345–352.

[209] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. "Learning transferable architectures for scalable image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8697–8710.