



HAL
open science

Multimodal deep learning for audiovisual production

Kaouther Ouenniche

► **To cite this version:**

Kaouther Ouenniche. Multimodal deep learning for audiovisual production. Machine Learning [stat.ML]. Institut Polytechnique de Paris, 2023. English. NNT : 2023IPPAS020 . tel-04480229

HAL Id: tel-04480229

<https://theses.hal.science/tel-04480229>

Submitted on 27 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS

NNT: 2023IPPAS020

Thèse de doctorat



Multimodal deep learning for audiovisual production

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom sud Paris

École doctorale n°626 Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat: Mathématiques et Informatique

Thèse présentée et soutenue à Evry, le 19/12/2023, par

Kaouther OUENNICHE

Composition du Jury :

Jenny BENOIS-PINEAU	Professeure, Université de Bordeaux (LABRI)	Présidente
Mohamed DAOUDI	Professeur, IMT Nord (CRIStAL)	Rapporteur
Amel BENZAZZA	Professeure, Sup'Com (COSIM)	Rapporteur
Andrei BURSUC	Docteur, Senior Researcher, VALEO.AI	Examineur
Titus ZAHARIA	Professeur, Télécom sud Paris (SAMOVAR)	Directeur de thèse
Ruxandra TAPU	Maître de conférences, Télécom sud Paris (SAMOVAR)	Co-Encadrante de thèse
Matthieu PARMENTIER	Senior Researcher, France TV	Invité



INSTITUT
POLYTECHNIQUE
DE PARIS

Acknowledgment

This Ph.D. journey has been a challenging and rewarding odyssey, and it is with immense gratitude that I recognize those who have been pivotal in making it possible.

First and foremost, I would like to express my deepest appreciation to my supervisor Pr. Titus Zaharia. His ideas and judgment have not just guided but profoundly inspired me throughout this demanding expedition. His advice and support have been a constant source of wisdom during these formative years. I would also like to thank Dr. Ruxandra Tapu, my co-Supervisor, who has been a wellspring of insightful commentary and encouragement when I found myself entangled in the complexities of my research. Her foresight and profound knowledge in my research area have consistently breathed life into my work, and for that, I am immensely grateful. My warm thanks also go to Mathieu Parmentier who brought a real-world perspective and invaluable insights, enriching the practical dimensions of my research. His guidance has expanded the horizons of my work, and I deeply appreciate his contributions.

I would like to extend my heartfelt appreciation to Pr. Mohamed DAOUDI, Pr. Amel BENZAÏZA, Pr. Jenny BENOIS PINEAU, and Dr. Andrei BURSU for graciously accepting to be members of the jury. Your willingness to share your feedback and valuable insights has been instrumental in shaping the outcome, and I am genuinely grateful for your dedication to advancing knowledge in this field.

I extend my heartfelt thanks to professors Catalin Fetita, Marius Preda, and Nicolas Rougon who engaged me in exciting conversations and provided insights that have enriched my academic and personal journey. I would also like to express my gratitude to Evelyne Taroni for her exceptional administrative support that has lightened the logistical burden and allowed me to focus on my work.

My academic pursuit would not have been the same without the camaraderie and friendship of my fellow colleagues in the ARTEMIS department. Abhaya-Dhathri Arige, Hugo Durchon, Antoine Didier, Traian Lavric, Zied Lahiani, Christian Tulvan, Minderis Krir, and Léa Saunier have shared this academic venture with me, offering not only their intellectual insights but also their friendliness and humor, which have made the journey all the more enjoyable.

To my beloved partner, Nicolas, I owe a debt of gratitude for your unyielding support and for pushing me beyond what I believed I was capable of. Your kindness and encouragement have been a constant source of strength and motivation.

Last but not least, I would like to express my heartfelt appreciation for my family and friends who have made my journey so much more meaningful. Even with miles that separate us, their unwavering presence in my life has been a constant source of comfort and strength. To my beloved nieces, your presence in my life has brought an abundance of joy and happiness. Your infectious laughter and zest for life have the power to light up even the darkest of days. To my dear sisters, you have delighted me in countless ways, offering a listening ear and a comforting shoulder to lean on when I have needed it most. To my parents, you hold a special place in my heart for making me the person I am today. Your love, guidance, and belief in me have been the driving force behind my growth and development. You have pushed me to be the best version of myself and instilled in me the values and principles that have guided my journey.

Table of contents

1	Introduction	1
1.1	Context	1
1.2	Experimental setup	2
1.3	Contributions	2
1.4	Thesis organization.....	3
2	Industrial use case	6
2.1	Introduction	7
2.2	Presentation of the study	7
2.2.1	In-Depth Interviews with Documentalists.....	8
2.2.2	Participation in the <i>Club des Archives de l’Audiovisuel Public</i> meetings.....	8
2.2.3	Participation to the European Broadcasting Union	9
2.3	Content management in the broadcast industry.....	10
2.3.1	Role of documentalists in content management	10
2.3.2	Television sources	11
2.3.3	Tools for content management and organization	12
2.3.4	Search process for efficient content discovery	12
2.3.5	Overview of metadata for indexing.....	14
2.3.6	Types of metadata used for indexing.....	15
2.4	Automation techniques.....	18
2.4.1	Automatic solutions to assist the documentalists in archive management	19
2.4.2	Proposed solutions.....	20
2.5	Conclusion.....	21
Part I: UNIMODAL, TASK-SPECIFIC models		22
3	Landmark recognition	23
3.1	Introduction	24
3.2	Related work.....	24
3.2.1	Challenges in landmark recognition.....	24
3.2.2	Content-based image retrieval.....	26
3.2.2.1	Traditional methods.....	26
3.2.2.2	Deep learning-based techniques	27
3.3	Constitution of a landmark dataset.....	28
3.4	Proposed methodology	31

3.4.1	Network architecture and image representation	31
3.4.2	Siamese learning.....	33
3.4.3	Dimensionality reduction and whitening.....	34
3.4.4	Image retrieval.....	35
3.5	Experiments and results.....	35
3.5.1	Datasets	36
3.5.2	Implementation details	36
3.5.3	Model evaluation on our dataset.....	36
3.5.4	Comparison with state-of-the-art.....	37
3.5.5	Qualitative results.....	37
3.6	Conclusion and future works	41
4	Contribution to the scene segmentation project	42
4.1	Introduction	43
4.2	Scene segmentation overview	43
4.3	Place recognition	44
4.3.1	Network architecture	44
4.3.2	Experimental setup	45
4.3.3	Model evaluation	45
4.4	Field of view shot detection	46
4.4.1	Experimental setup	47
4.4.2	Model evaluation.....	48
4.5	Conclusion.....	50
5	Camera motion categorization.....	51
5.1	Introduction	52
5.2	Types of camera motion	53
5.3	Related work.....	54
5.3.1	Estimation of motion vectors.....	54
5.3.1.1	Feature-based approaches.....	54
5.3.1.2	Appearance-based approaches.....	54
5.3.1.3	Discussion	55
5.3.2	CNNs for video action recognition.....	55
5.3.2.1	Two-stream networks	56
5.3.2.2	3D Convolutional Neural Networks	57
5.3.2.3	Discussion	58
5.4	Proposed methodology	58
5.4.1	Network architecture	58
5.4.2	Camera motion datasets.....	60

5.4.2.1	Semi-automatic learning dataset creation.....	60
5.4.2.2	Creation of the test dataset	68
5.5	Experimental results	70
5.6	Conclusion.....	73
Part II: Multimodal models		74
6	Multimodal learning.....	75
6.1	Introduction	76
6.2	Applications of multimodal learning for TV broadcast.....	76
6.3	Challenges in multimodal learning.....	77
6.3.1	Data heterogeneity.....	77
6.3.2	Data fusion	77
6.3.3	Alignment.....	78
6.3.4	Efficiency	78
6.4	Transformer architecture for multimodal learning	79
6.4.1	Mathematical formulation of the Vanilla transformer.....	79
6.4.1.1	Multi-Head Self Attention.....	80
6.4.1.2	Position-wise Feed-Forward Networks	80
6.4.2	Advantages of the transformer architecture.....	82
6.4.3	Challenges in transformers	82
6.5	Landscape of multimodal datasets.....	83
6.6	Conclusion.....	84
7	Video question answering	86
7.1	Introduction	87
7.2	Application of Video Question Answering for archive indexing and retrieval.....	88
7.3	Related work.....	88
7.3.1	VideoQA datasets and evaluations metrics	88
7.3.1.1	VideoQA datasets.....	88
7.3.1.2	Evaluation metrics	91
7.3.1.3	Discussion	92
7.3.2	State of the art VQA techniques	93
7.3.2.1	Monolithic models with attention.....	94
7.3.2.2	Memory-based models	94
7.3.2.3	Graph-based models	95
7.3.2.4	Transformer-based models	95
7.3.3	Discussion	98
7.4	Proposed network architecture	99
7.4.1	Feature extraction	100

7.4.1.1	Video processing	100
7.4.1.2	Text processing.....	101
7.4.2	Cross-modal module.....	102
7.4.3	Transformer-based multimodal fusion	103
7.5	Rephrasing attacks.....	104
7.5.1	Problem formulation.....	104
7.5.2	Methodology	105
7.6	Experimental evaluation.....	107
7.6.1	Datasets	107
7.6.2	Implementation details	107
7.6.3	Ablation studies	108
7.6.3.1	Ablation studies on MSVD-QA	108
7.6.3.2	Effect of the transcript input.....	111
7.6.4	Comparison with state-of-the-art.....	113
7.7	Conclusion.....	113
8	Video captioning	115
8.1	Introduction	116
8.2	Application to TV archive indexing	117
8.3	Related work.....	117
8.3.1	Template-based approaches.....	117
8.3.2	Deep-learning based approaches	118
8.3.2.1	Visual-based approaches	119
8.3.2.2	Multimodal approaches	119
8.3.3	Evaluation metrics	122
8.3.3.1	BiLingual Evaluation Understudy (BLEU).....	122
8.3.3.2	Recall-Oriented Understudy for Gisting Evaluation (ROUGE).....	122
8.3.3.3	Metric for Evaluation of Translation with Explicit ORdering (METEOR).....	123
8.4	Proposed video captioning architecture.....	124
8.4.1	Feature extraction	124
8.4.1.1	Visual feature representation	124
8.4.1.2	Textual feature representation	125
8.4.2	Modality Attention module	126
8.4.3	Transformer encoder	128
8.4.4	Transformer decoder	129
8.5	Training objectives	132
8.5.1	Masked Language Modeling	132
8.5.2	Contrastive learning.....	133

8.5.3	Caption generation.....	133
8.6	Experiments and results.....	133
8.6.1	Dataset	134
8.6.2	Implementation details	135
8.6.3	Ablation study	135
8.6.4	Comparison with state of the art.....	137
8.6.5	Qualitative results.....	138
8.7	Conclusion.....	142
9	Conclusion and perspectives	144
9.1	Conclusion.....	144
9.2	Future work	145
	References	148



List of figures

Figure 2.1. Distribution of the number of participants in the meeting with respect to their institution affiliation.	9
Figure 2.2. Main activities of the documentalists.	11
Figure 2.3. Overview of the DALET tool for content organization and management.	13
Figure 2.4. Top search items conducted by documentalists.	13
Figure 2.5. Top actions researched by documentalists.	14
Figure 2.6. Example of an archive file. Source: DALET.	16
Figure 2.7. Example of tagged-keywords in an archive file. Source: DALET.	17
Figure 2.8. Index page for election candidates. Source: ELEC+.	18
Figure 3.1. Variations in viewpoint, illumination and presence of distractors.	25
Figure 3.2. Three look-alike gothic churches. (a) Notre Dame, (b). Amiens, (c). Reims.	25
Figure 3.3. Overview of the image retrieval framework.	26
Figure 3.4. Siamese network architecture with contrastive loss.	28
Figure 3.5. Examples from users' pictures falsely tagged as a landmark. (a) Eiffel tower (b) Arch of Constantine (c) Empire state building (d) Statue of liberty.	29
Figure 3.6. Examples from the dataset.	30
Figure 3.7. Network training using contrastive loss (offline).	32
Figure 3.8. Feature aggregation using MAC technique.	32
Figure 3.9. Similarity learning task. The objective is to minimize the distance between positive samples and to maximize the distance between negative samples.	33
Figure 3.10. Example of batch-wise positive/negative mining.	34
Figure 3.11. Overview of the image retrieval process.	35
Figure 3.12. Retrieval examples from our dataset.	38
Figure 3.13. Retrieval examples from Paris6k dataset.	40
Figure 4.1. The AI-TV ads insertion framework.	43
Figure 4.2. Inception block.	44
Figure 4.3. Examples of the top-5 predictions from France TV content. The number beside indicates the prediction confidence.	46
Figure 4.4. Basic field of view shot types. a) EWS; b) LS; c) MS; d) MCU; e) CU; f) ECU.	47
Figure 4.5. Confusion matrix of the test set.	48
Figure 4.6. Examples of field of view shot recognition on France TV content. Horizontal bars indicate the prediction confidence.	48
Figure 4.7. Examples of Field of view shot type prediction. (GT: Ground Truth, P: Prediction).	49
Figure 5.1. Different types of camera movement.	53
Figure 5.2. Workflow of two-stream network [64].	56
Figure 5.3. Workflow of 3D CNN.	57
Figure 5.4. Illustration of the adopted 3D CNN. The notation $F@H^3$ means F filters of size $H \times H \times H$	59
Figure 5.5. Skip connection employed in the network.	59
Figure 5.6. Examples from the videos of the dataset. (a) Pan Left, (b) Tilt Up, (c) Static.	61
Figure 5.7. Grid of points in a frame.	61
Figure 5.8. Correspondence between interest points in two successive frames.	62
Figure 5.9. Dominant angle and distance across four regions in a frame. (a) Pan Right; (b) Pan Left; (c) Tilt-up; (d) Tilt Down; (e) Zoom In; (d) Zoom-out; (g) Static; and (h) Unknown.	65
Figure 5.10. Examples from the Training dataset.	67
Figure 5.11. Examples from the test dataset. The videos includes high-resolution samples as well as hand-shake videos, blurry images and illumination variations.	69

Figure 5.12. Comparison of the different configurations: (a). The loss variation (b). The accuracy variation. (Blue: Resnet trained from scratch, Red: Resnet + reverse frames, Orange: Resnet + finetuning, Green: Resnet + finetuning + reverse frames)	70
Figure 5.13. Confusion matrix of the validation dataset	71
Figure 5.14. Examples of recognition results on the test dataset.....	72
Figure 6.1. Architecture of the vanilla transformer [4].	81
Figure 7.1. Video Question Answering task.	87
Figure 7.2. Examples from different datasets. (a) MSRVTT-QA; (b) ActivityNet-QA; (c) KnowIT; (d) SocialIQ; (e) CLEVER; (f) MSVD-QA.	90
Figure 7.3. Basic VideoQA Framework.	94
Figure 7.4. Task-agnostic training paradigm.....	96
Figure 7.5. The proposed framework for Video Question Answering task.....	99
Figure 7.6. The video transformer architecture.	100
Figure 7.7. Overview of text processing framework.	101
Figure 7.8. Cross-modal correlation module.....	103
Figure 7.9. Rephrasing attacks on Video Question Answering model.....	106
Figure 7.10. Examples of results of our approach on the MSVD-QA dataset, with both original and rephrased questions. OQ: Original question; RQ: Rephrased Question; GT: Ground Truth; OP: Prediction of the model to the Original question; and RP: Prediction of the model to the rephrased question.	110
Figure 7.11. Examples of predictions on MSRVTT-QA dataset.	112
Figure 8.1. Video captioning problem.....	116
Figure 8.2. An archive indexing page. The field « descriptif » represents the natural language description of the video content at the shot level. Source: DALET.....	117
Figure 8.3. The paradigm of the encoder-decoder architecture.....	118
Figure 8.4. Various paradigms for video-text training. (a) Share-type; (b) Cross-type; and (c) Joint-type.	121
Figure 8.5. Overview of the proposed multi-modal architecture.	124
Figure 8.6. The modality attention module.	126
Figure 8.7. Video samples from MSRVTT dataset for which the transcript and video data are not well-aligned.	127
Figure 8.8. Overview of the encoder architecture. (left) Encoder block. (right) multi-head self-attention mechanism.....	129
Figure 8.9. Overview of the decoder architecture.	130
Figure 8.10. Multi-head Cross Attention process.....	131
Figure 8.11. (a) Sample requiring both transcript and visual modalities for caption generation. (b) Sample requiring visual cues only.....	134
Figure 8.12. Qualitative results from MSRVTT dataset. Samples requiring both textual and visual modalities to generate the caption.	140
Figure 8.13. Qualitative results from MSRVTT dataset. The ASR is not aligned with the content of the video.	141
Figure 8.14. Qualitative results from MSRVTT dataset. Samples with no audio channel.....	142

List of tables

Table 2.1. Departments and people involved in the considered industrial case study at France TV.	8
Table 3.1. Geo-coordinates of five landmarks.	31
Table 3.2. Comparison with state-of-the-art methods in landmark recognition and retrieval tasks on Paris6k dataset.	37
Table 5.1. The number of videos in each category in the train/val dataset	66
Table 5.2. The number of videos per category in the test dataset	68
Table 7.1. Statistics of VideoQA datasets.	91
Table 7.2. Examples of rephrased questions from MSVD-QA dataset.	107
Table 7.3. Ablation studies on MSVD-QA. Acc1 represents the performance on the original dataset. Acc2 represents the performance on the rephrased dataset.	109
Table 7.4. Comparison of the effect if the transcript input on MSRVTT-QA dataset.	111
Table 7.5. Comparison with state-of-the-art models on MSVD-QA and MSRVTT-QA	113
Table 8.1. Ablation studies on MSRVTT dataset.	135
Table 8.2. Performance comparison (BLEU4) across models using different input modalities on two subsets	136
Table 8.3. Statistics of video captioning models. PT stands for Pre-Training. x stands for unknown	137
Table 8.4. Comparison with state of the art.	138

1 INTRODUCTION

1.1 Context

Television has long been a primary source of information and entertainment, shaping our understanding of the world and capturing significant moments in history. With the proliferation of television channels and the vast amount of content generated daily, the need for effective indexing and organization of television archives has become increasingly critical. This is the case for prominent broadcasters such as France Television (France TV), a leading French television network. France TV produces a diverse array of television content across a great variety of genres. From news programs that provide vital information to game shows that entertain audiences and television series that captivate viewers, the breadth and depth of content are vast. Managing such a diverse range of content presents unique challenges for France Television in effectively organizing and indexing their huge television archive. Each genre and format require specific considerations in terms of indexing criteria, metadata extraction, and retrieval methods.

Within this framework, documentalists play a crucial role in the management of TV's archive, ensuring the indexing and retrieval processes are carried out efficiently. Their primary responsibility is to index the diverse range of content, making it easily searchable and reusable for future purposes. Documentalists meticulously analyze the content, assigning relevant keywords and crafting natural language descriptions to capture the essence of each program. Once indexed, the data is sent to the Institut National de l'Audiovisuel (INA), where it is stored and preserved. When it comes to retrieving the data, documentalists rely on the keywords and natural language descriptions they previously tagged. It is crucial for the data to be well indexed, as this ensures effective retrieval and maximizes the reusability and accessibility of the content for researchers, journalists and general public. The diligence and expertise of documentalists in the indexing and retrieval process are pivotal in unlocking the value of France TV's archive and enabling seamless access to its rich and diverse content.

However, documentalists are often faced with limited resources and tight deadlines, leaving them with insufficient time to dedicate to the meticulous process of indexing. As a result, there is a greater emphasis on information retrieval rather than comprehensive indexing. Their primary focus becomes locating specific content requested by researchers or fulfilling immediate needs, leaving little room for the thorough indexing necessary to maximize the archives' long-term utility.

Our doctoral research has been carried out within the framework of the AITV (Artificial Intelligence for TV) joint laboratory, established in 2019 between France TV and Télécom Sud Paris. AI systems can automatically generate comprehensive metadata, enabling documentalists to shift their focus towards more critical tasks, such as curating and verifying the quality and relevance of information. Our objective is to leverage AI methods to alleviate the burden on documentalists and enhance the efficiency of their work.

To achieve our objective, we have first performed an in-depth analysis of their job responsibilities and challenges. We have initiated a series of meetings and discussions with documentalists from various departments within France TV. Through these interactions, we have gained valuable insights into the intricacies of their daily tasks and the difficulties they encounter. This collaborative approach allowed us to foster a deep understanding of their specific needs. Based on this knowledge, we have engaged a joint brainstorming process to devise AI-empowered solutions specifically tailored to their requirements. The proposed solutions are not static but actively tested and refined based on continuous feedback from the documentalists. By aligning our work with their input, we aimed to develop practical tools and approaches that can effectively address their challenges and significantly improve their workflow.

In our research, we have undertaken a comprehensive analysis of the needs and research patterns commonly encountered by documentalists. We have identified several key topics that are manually indexed in the television audio-visual archive. Our ultimate goal is to develop a robust framework for multi-modal indexing of such content, with the help of deep learning techniques. In an initial phase of our work, we have focused on implementing task-specific, unimodal models which identify keywords and tags currently handled manually by documentalists. Such models encompass landmark recognition, place recognition, field of view shot type identification, and camera motion categorization. In the subsequent phase of our research, we have expanded our approach to incorporate comprehensive semantic analysis of the video through the examination of different modalities. To this purpose, we have developed various models for video question answering and video captioning, which provide a more holistic understanding of the content.

1.2 Experimental setup

The experimental setup for this research encompassed two distinct phases, each requiring specific hardware configurations.

The initial phase of this Ph.D. research, focusing on task-specific unimodal models, has been conducted on the "Thanos" server located in our laboratory. This server is equipped with two NVIDIA GTX 1080 GPUs.

The subsequent phase of the PhD research was centered around the development of multimodal models. To meet the increased computational demands of combining information from multiple modalities effectively, we utilized a separate hardware configuration. Specifically, this phase was carried out on a system equipped with two NVIDIA GeForce RTX 2080 GPUs. The choice of these GPUs for multimodal learning was driven by the need for more powerful hardware to accommodate the increased complexity of processing multiple types of data simultaneously. Multimodal models involve the fusion of information from sources such as text, images, and audio, necessitating GPUs with greater computational capacity to expedite the convergence of these intricate models.

Irrespective of the hardware infrastructure used, all models in this research have been developed using Python3 programming language and the PyTorch framework. Our selection of the PyTorch framework for this research is rooted in its exceptional flexibility, vibrant ecosystem, and strong community support. PyTorch's dynamic computational graph offers us a crucial advantage by simplifying the development and modification of intricate neural network architectures, allowing us to adapt our models rapidly as needed. Python's versatility complements PyTorch seamlessly, enabling us to harness a vast array of open-source libraries and tools for deep learning, data manipulation, and visualization. This rich ecosystem empowers us to explore and integrate cutting-edge techniques and pre-existing solutions into our research pipeline efficiently. Furthermore, the extensive documentation and active community surrounding PyTorch provide valuable resources for troubleshooting, optimization, and knowledge-sharing, enhancing the overall robustness of our work.

1.3 Contributions

This thesis focuses on leveraging deep learning techniques to improve the indexing of television archives. It deals with the generation of various keywords and tags from audiovisual archives and delves into the exploration of research challenges that can be approached through both unimodal and multimodal strategies. Our research was tested on television content as well as standard benchmark datasets, and we have validated its performance against previous state-of-the-art methods.

The first contribution concerns the classification of shots according to the camera motion type. We propose the first data-driven solution to address the problem. Specifically, we introduce a novel approach based on 3D convolutional neural networks with residual blocks, inspired by action recognition techniques. We apply transfer-learning technique to overcome the data scarcity issue for camera motion characterization. We initially train our model on the Kinetics [38] action recognition dataset. The Kinetics corpus has nothing to do with our purpose. However, we claim that the derived feature maps capture essential, salient spatio-temporal cues that can be exploited for our task. A fine-tuning is then applied on a dedicated camera motion data set with a reduced number of items. Additionally, we propose a semi-automatic method that makes it possible to construct a reliable camera motion dataset from general public videos with a minimum amount of human intervention. Finally, the third contribution concerns the creation of a camera motion evaluation dataset. The corpus includes highly challenging videos, acquired in real-life conditions with professional cameras and at various resolutions. It allowed us to assess the robustness and power of generalization of the proposed technique, which yields an average accuracy rate of about 94%.

The second contribution addresses the challenging task of Video Question Answering (VideoQA). Our objective is to explore the power of attention-based transformers, which have proven highly effective in natural language processing, for the purpose of grounded multimodal learning in the context of VideoQA. We first point out the challenges in developing an efficient, computationally feasible model due to the inherent differences between visual and textual modalities and the quadratic complexity of transformers. To overcome such limitations, a novel framework incorporating a lightweight transformer in conjunction with a cross-modality module is proposed. The latter uses cross-correlation to facilitate the reciprocal learning of text-conditioned visual features and video-conditioned textual features. To assess our model's robustness and real-world applicability, we introduce an adversarial testing scenario with rephrased questions. Additionally, we investigate the significance of the transcript modality in predicting accurate answers, conducting ablation studies on the MSRVTT-QA dataset [1], including subsets that require only visual information or both the transcript and video for answer generation. This comprehensive work offers insights into the vulnerability of VideoQA models to linguistic variations and the importance of the transcript modality. The experimental evaluation, carried out on the MSVD-QA and MSRVTT-QA benchmark datasets, validates the proposed methodology with average accuracy scores of 44.96% and 41.88% respectively. When compared with state-of-the-art methods the proposed method yields gains in accuracy of more than 2%.

The third contribution tackles the issue of multimodal video captioning. We introduce a novel framework including a modality-attention module that captures the relationships between visual and textual data using cross-correlation. Additionally, we integrate temporal attention to extract contextual information from a 3D CNN, enhancing the model's ability to produce meaningful captions. Notably, our work introduces an auxiliary task utilizing a contrastive loss function, promoting the generalization of the model and a deeper understanding of inter-modal relationships and underlying semantics. By comparing the multimodal representation of the video-transcript with the caption representation, we achieve improved performance and ensure knowledge transfer. The utilization of a transformer architecture for encoding and decoding effectively captures interdependencies between text and video information through attention mechanisms. The experimental evaluation, carried out on the MSRVTT benchmark [2], validates the proposed methodology, which achieves BLEU4, ROUGE, and METEOR scores of 0.4408, 0.6291 and 0.3082, respectively. When compared to the state-of-the-art methods, the proposed approach shows superior performance, with gains in performance ranging from 1.21% to 1.52% across the three metrics considered.

1.4 Thesis organization

The rest of the manuscript is organized as follows.

In Chapter 2, we lay out the context and objectives of our work. This initial part explores the intricate process of archive indexing and retrieval within France TV. A significant attention is given to the critical role held by documentalists in this system, acknowledging the challenges they encounter in their daily responsibilities. The rationale underpinning the selection and design of specific solutions is thoroughly discussed.

The rest of the manuscript is divided into two parts.

Part I includes three chapters and is devoted to the design and development of task-specific, unimodal models. Such models are primarily engineered to label the variety of key metadata commonly found in an archive file.

Chapter 3 provides a thorough examination of the issue of landmark recognition. As a crucial facet of French cultural heritage, landmarks feature prominently in archival indexing. The extensive literature dedicated to this subject is first presented and analyzed. The review of the state-of-the-art shows that existing publically available landmark datasets are not well-aligned to the specific requirements of France TV. However, the availability of an adapted landmark dataset that can be used for learning/training objectives within the recognition process, and that can cover landmarks of interest for France TV is a crucial prerequisite to the design of a successful landmark recognition solution. Consequently, we have first proposed an automatic dataset construction methodology that can be notably customized for the France TV requirements. This dataset, in our case, includes a set of significant landmarks listed in the archives. Subsequently, a landmark recognition method is proposed. We have adopted a recognition-by-retrieval approach, which offers the advantage of being able to cope with the dynamic, time-varying character of the landmark thesaurus, without needing any model retraining following each update of the considered thesaurus. The underlying similarity measure is obtained with the help of a Siamese network with contrastive loss. A dimensionality reduction technique is also applied to optimize the model's efficiency. The performance of the proposed model is assessed in a zero-shot setting, both on the custom France TV dataset and on the Paris6k benchmark [3].

In chapter 4, we present our contribution to a larger project, carried out within the joint IATV laboratory, which concerns the automatic scene segmentation of TV content. The scene segmentation process is based on a shot clustering approach, and exploits the similarity between various visual components. Our contributions concern the automatic identification of visual cues that can be further exploited by the scene segmentation process. More precisely, we have considered place recognition (e.g. indoor, outdoor, studio, restaurant, park) and field of view shot type identification methodologies. The performances of the proposed solutions have been experimentally evaluated on both public datasets and on a FranceTV benchmark that we have specifically created to this purpose.

Chapter 5 shifts focus to the issue of estimating the camera motion's type. The camera motion's type is an important feature, associated to individual shots and specified by documentalists during the indexing process. The literature review shows that prevalent state-of-the-art methods often rely on traditional techniques such as interest point estimation and tracking for predicting camera motion. In contrast, we propose a novel deep-learning based methodology, inspired from recent advancements in the field of action recognition, and based on a 3D CNN (*Convolutional Neural Network*) model specifically designed to this purpose. As the available datasets were not suitable for training our model, we have developed two distinct datasets. The first one involves a collection of random videos from YouTube, which have been indexed through automated processes derived from traditional techniques. In order to fully assess the validity of our approach, we have constructed a second dataset manually, employing multiple cameras of different types and characteristics. This second dataset presents significant challenges, including variations in illumination, hand-shake disturbances, diversity in resolution and frame rates. The experimental results obtained on both datasets demonstrate the pertinence of the proposed approach, with recognition rates of 97% and 94% on the first and second datasets, respectively.

Part II of the manuscript delves into the exploration of multimodal models, which require the integration and fusion of various modalities, including video, text, image, and audio.

An extensive state-of-the-art survey is first proposed (Chapter 6). Here, we present the key, pervasive challenges in multimodal learning such as data heterogeneity, fusion, alignment and scalability. Subsequently, we present the most recent and promising approach dedicated to this task, which is the transformer model [4]. We describe the mathematical formulation of the transformer and discuss its advantages and challenges in comparison to previous state-of-the-art models. The chapter concludes with a presentation of existing multimodal datasets, which are crucial for successful learning, and their progression over time.

Chapter 7 introduces a first multimodal technique, which concerns a novel Video Question Answering (VideoQA) methodology. VideoQA consists in providing a coherent answer to a question related to the content of a given video. A comprehensive review of the state of the art is first presented, with identification of main families of methods and analysis of related strengths and limitations. Subsequently, the proposed methodology is described in details. The originality of the methods comes from the joint integration of a multimodal transformer model, conceived to solve the fusion problem, and of a cross-modal model, designed to address the alignment issue. In particular, the approach makes it possible to set up a lightweight transformer model, compatible with TV-related applications. The vulnerability of the composing elements of our pipeline is tested using black box attacks that represent automatically-generated, semantic-preserving rephrased questions. We demonstrate through ablation studies the effectiveness of each block in our framework to improve the performance and generalization of our approach. The experimental evaluation, carried out on the MSVD-QA and MSRVTT-QA benchmark datasets [1], validates the proposed methodology with average accuracy scores of 44.96% and 41.88% respectively. When compared with state-of-the-art methods the proposed method yields gains in accuracy of more than 2%.

In natural continuity with the related VideoQA developments, Chapter 8 proposes a new multimodal video captioning method. The video captioning process aims at providing a short summary expressed in natural language that semantically summarizes the video content. A state of the art review is first presented. Let us underline that video captioning methodologies are closely related to VideoQA in many aspects. The major difference concerns the decoder part. For VideoQA, in most of the cases, it consists of a classification head implemented on a vocabulary of possible answers. For video captioning, the decoder is based on RNN or transformer architectures. The proposed network architecture is then detailed. We have considered a transformer model that takes as input both the visual and the subtitle textual representations. Similarly to the previous VideoQA approach introduced in Chapter 7, a cross-modal module is also integrated here for modality alignment purposes. An additional contrastive loss function is introduced in order to optimize, during the learning process, the alignment between the video-text multimodal representation and the corresponding caption. Finally, the caption is generated using a transformer decoder with a teacher-forcing method. The proposed approach is validated through various ablation studies and comparisons with recent state-of-the-art methods.

Finally, Chapter 9 concludes the manuscript and summarizes the main contributions of this work. We also outline potential future research directions in terms of methodologies and related applications.

2 INDUSTRIAL USE CASE

Abstract: In this chapter, we provide an overview of the context and objectives of our research. The foundation of our work is based on a series of interviews and meetings conducted with various stakeholders in the broadcast community. These interactions have provided valuable insights into the work of documentalists and the challenges they encounter in their daily operations.

We start the chapter by presenting the details of our research study and our collaboration with France TV. We delve into the content management methods employed by France TV for their archives, highlighting their significance in the overall workflow. Furthermore, we emphasize the pivotal role played by documentalists in the process of indexing and highlight various types of metadata associated with the archived content. Next, we shed light on the complexities and intricacies involved in manual indexing, discussing the challenges documentalists face in this regard. Subsequently, we delve into the potential of automation techniques to enhance and expedite the indexing process while ensuring standardization across the board. In the final part of this chapter, we elaborate on the solutions proposed in this PhD research and we highlight the specific reasons that motivated their selection

Keywords: Archive indexing, TV broadcast, documentalists, deep-learning.

2.1 Introduction

We are amidst a technological revolution powered by advances in Artificial Intelligence (AI), which is reshaping numerous industries, including media and entertainment. France TV, along with other broadcasters, is racing to integrate advanced technologies to stay competitive, profitable, and meet market needs. The goal is to speed up product development, reduce resource consumption, and improve design quality. One crucial area where AI can make a substantial impact is related to the management of audio-visual archives.

The manual process of indexing such huge archives is inefficient and time-consuming. Documentalists traditionally perform indexing manually, spending countless hours on this tedious task. This is where our PhD research project comes in - to explore AI applications for improving archive management in TV broadcasting.

This chapter provides an in-depth look into the initial PhD phase. In section 2.2, we detail the study conducted to establish the objectives of this PhD project. Our research approach included extensive engagement with various stakeholders from the broadcast community. We have conducted numerous interactions with documentalists working across different departments and regions. These interactions provided us with a thorough understanding of the present state of the archive management processes. Further enriching our perspective, we attended meetings with other broadcasters who are already leveraging AI solutions. In these forums, they shared their plans for future developments aimed at refining their broadcasting capabilities. This exposure offered us valuable insights into the possibilities and potential of AI in the broadcast industry, which significantly oriented our research directions and objectives.

In the section 2.3, we examine the current state of content management in France TV. A particular emphasis is put on the indexing process, which prompted a detailed exploration of various sources that require indexing, including news segments, pool materials, and other relevant sources. Moreover, we delve into the specific indexing tools employed by documentalists, such as Dalet, and investigate the methodologies currently employed in the indexing process.

The final section of this chapter discusses the challenges faced by these documentalists and the potential AI-powered solutions that can help overcome these issues. We present the AI solutions chosen for our research, with the rationale behind their selection.

2.2 Presentation of the study

The primary aim of the study was to gain a comprehensive understanding of the pivotal role played by documentalists at FranceTV and their significant contribution to the indexing and research process of the archives. This encompassed examining their involvement in meticulously organizing and categorizing the vast collection of archival materials, ensuring easy accessibility and retrieval of information. Furthermore, the study aimed to uncover the nuances of the indexing process employed by documentalists, shedding light on the techniques, methodologies, and tools utilized in their work. We sought to identify the challenges faced by documentalists, such as the time and effort required for manual indexing and the potential limitations or inefficiencies within the existing system. Moreover, the study aimed to explore the crucial relationship between documentalists and the broader broadcast community at FranceTV. This involved examining how documentalists collaborate with researchers, journalists, and other stakeholders to fulfill their information needs and support the production of high-quality content. The research aimed to uncover the dynamics of this collaboration, identifying any areas where improvements or synergies could be fostered. By gaining a comprehensive understanding of the documentalists' role, we aimed to identify potential areas for improvement and the application of

emerging technologies, such as artificial intelligence and machine learning. The ultimate objective was to enhance the efficiency, accuracy, and accessibility of the archive.

2.2.1 In-Depth Interviews with Documentalists

As a part of our research, we have conducted several in-depth interviews with documentalists from various departments within FranceTV, including News (at both national and regional levels), Overseas, Sports, and Politics. The details of the interviewees are presented in Table 2.1.

Table 2.1. Departments and people involved in the considered industrial case study at France TV.

Department	Position
News (National)	Documentalist and Manager
Politics	Documentalist
News (Regional)	Documentalist
News (Overseas)	two documentalists
Sports	Manager and team of documentalists

The interviews have been conducted by a team comprising of myself, a Product Owner from the DAIA (Data and AI) department of France TV, and an AI Engineer from France TV.

Each interview session lasted approximately three hours, providing ample time for comprehensive discussion and exploration. During these sessions, documentalists thoroughly explained the methodology of their work, demonstrated the tools they use, and expounded upon the metadata they index. These interactions offered us the opportunity to understand a typical day in the life of a documentalist at FranceTV, allowing for deeper insights into their responsibilities, tasks, and processes.

Toward the end of each interview, we facilitated a discussion focusing on the challenges documentalists encounter in their daily work. These challenges were either openly expressed by the documentalists or surfaced through probing questions from the interviewers. We also conducted brainstorming sessions to generate ideas for potential improvements using AI tools.

These interviews significantly enriched our understanding of the documentalists' role, their operational processes, the challenges they face, and the possible areas where the application of AI could enhance their work efficiency and output quality. The findings from these interviews have greatly contributed to the study, providing key insights that would help us in formulating AI-based solutions for archive indexing and information retrieval.

2.2.2 Participation in the *Club des Archives de l'Audiovisuel* Public meetings

Throughout the course of our research, we had the opportunity to actively participate in five meetings of the Club des Archives de l'Audiovisuel Public. These meetings, organized by the Institut National de l'Audio-visuel (INA), encompassed documentalists from various public broadcasters, providing a rich and diverse forum for discussion. The documentalists involved in these meetings represented a broad spectrum of institutions, including INA, Radio France, France TV, ARTE, and France 24. Figure 2.1 provides a detailed list of the number of participants.

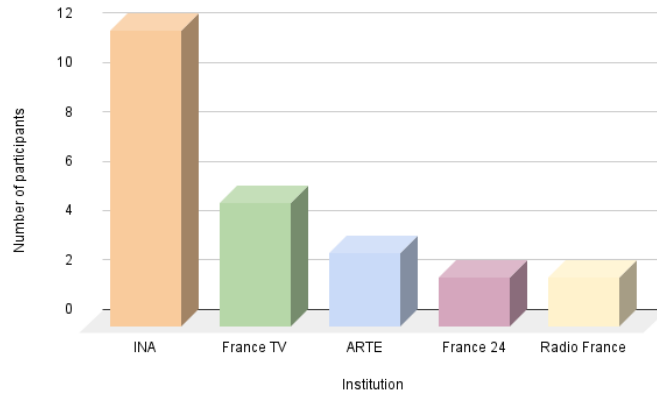


Figure 2.1. Distribution of the number of participants in the meeting with respect to their institution affiliation.

The interviewing team was also present during these meetings, further strengthening the communication and interaction between the researchers and the documentalists. The primary objective of these gatherings was to discuss the evolution of the documentalist job within public broadcasters, and the introduction and adaptation of new tools and innovative ideas. From our research perspective, these meetings were instrumental in sharing the progress of our PhD project and receiving valuable feedback from industry experts. This interaction enabled us to refine and fine-tune our AI-based solutions in response to the constructive feedback received.

As a part of our data collection strategy, we distributed a survey form among the documentalists participating in these meetings, requesting them to share it with their respective departmental colleagues. The survey, which resulted in 31 responses, aimed to assess the documentalists' exposure to AI tools, understand the frequent research tasks they perform, and the metadata they index. These insights would help prioritize our projects by aligning them more closely with the needs and challenges faced by documentalists.

This approach provided us with a wealth of data that significantly contributed to our research. A detailed discussion of the key findings from the survey and their implications will be presented in the subsequent sections of this chapter.

2.2.3 Participation to the European Broadcasting Union

The European Broadcasting Union (EBU) is the world's preeminent alliance of public service media, incorporating 112 member organizations across 56 countries, and overseeing nearly 2,000 television, radio, and online channels and services. As part of our research, we actively participated in one of the EBU meetings, designed to explore the ongoing impact of artificial intelligence on the broadcasting business.

During the meeting, several teams showcased AI-based solutions intended to augment various aspects of the audiovisual production. The goal was to illustrate the application of AI technology in optimizing diverse areas of broadcasting such as content creation, scheduling, data analysis, archive management and audience engagement.

Apart from sharing our AI-based solutions, the EBU meeting provided an invaluable opportunity for brainstorming. Collective discussions, spurred by a spirit of collaboration and innovation, allowed all participants to share insights and potential AI strategies aimed at enhancing audiovisual production.

Three key outcomes resulted from our participation in the EBU meeting. Firstly, we received vital feedback on our AI solutions from a variety of industrial experts. This feedback has significantly assisted us in refining our tools to better align them with the needs of the broadcasting community. Secondly, exposure to other AI initiatives and developments within various broadcasting organizations offered insight into the current state of AI application in the industry, guiding our research towards more innovative AI tools and techniques. Lastly, the rich discussions and brainstorming sessions during the meeting sparked a plethora of ideas for future research projects, steering our research trajectory towards further exploration of AI's potential in transforming the broadcast industry.

The insights derived from this study significantly influenced the trajectory of our research project, helping to articulate the primary goals and pinpoint the essential research areas to delve into. Let us now discuss these key findings in detail.

2.3 Content management in the broadcast industry

In this section, we will delve into the significant findings drawn from our collaborative interactions as outlined in section 2.2. We notably detail the process of content management within France TV, an integral understanding of which underpins the motivations driving our research.

2.3.1 Role of documentalists in content management

The documentalists are mainly responsible for the indexing of programs once they are transmitted. They review the subjects and provide necessary comments to enrich the audiovisual heritage and reuse the metadata, as requested by journalists. They are divided by territory and themes, with a focus on news and sports. Documentalists have an global view on the content and can anticipate or propose ideas for creating a topic. They can actively monitor identified or yet-to-be-identified subjects. The media library at the headquarters in Paris consists of approximately 40 documentalists distributed across five working units: central research, indexing, editorial services, photo library, and print media. The allocation of activities within the media library is roughly as follows: Indexing - 50% of the activity and Search and retrieval - 50% of the activity. As illustrated in Figure 2.2, the main activities of a documentalist include:

- Enriching the documentary collections by selecting, indexing, and archiving documents.
- Conducting document searches (both internal and external sources) for users and providing necessary documents (images, sounds) for content production (programs, shows, web publications, etc.).
- Proposing and creating structured documentary files.
- Associating copyright and distribution-related information with the documentary collections.
- Performing ongoing documentary monitoring in their field of activity, particularly in the digital realm (e.g., Facebook, internet). The documentalists maintain constant digital monitoring throughout the day, including regional news websites, Twitter, Facebook, and other social media platforms.

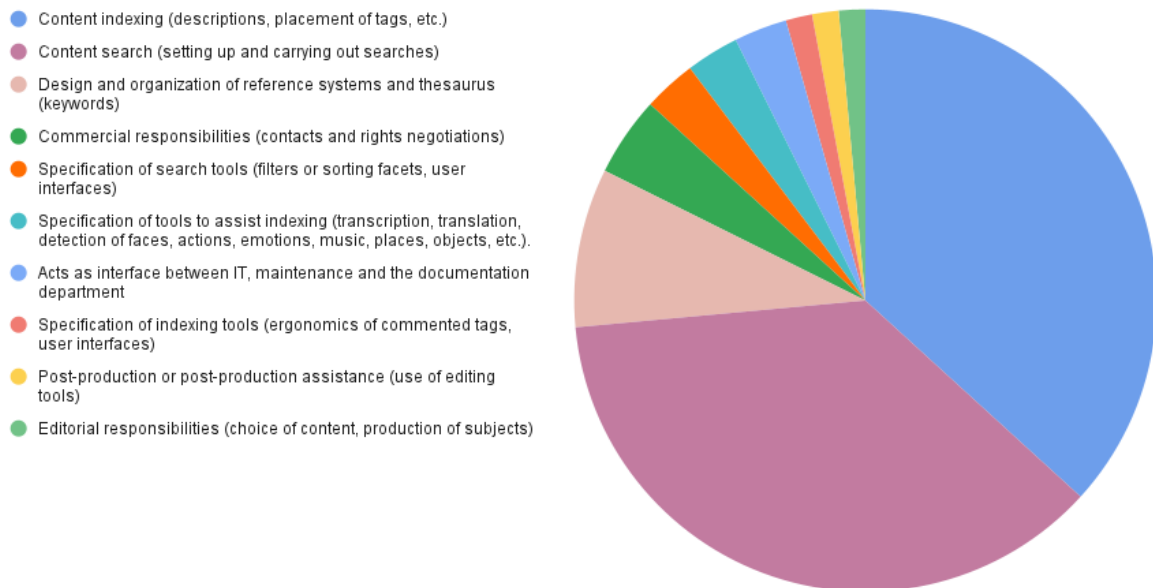


Figure 2.2. Main activities of the documentalists.

In our study, we have identified significant potential for AI solutions in the indexing process, which is particularly relevant considering the time-consuming nature of indexing. The primary objective for documentalists is to assist journalists in efficiently locating the relevant content through effective search methods, rather than dedicating excessive time to the indexing process. Therefore, our research emphasizes the need to streamline and automate indexing tasks using AI technologies, allowing documentalists to focus on their essential role of facilitating content discovery for journalists.

2.3.2 Television sources

In the process of indexing, various sources are to be considered for proper organization and accessibility. In terms of video content, the sources to be indexed include subjects broadcasted on the France TV network, content aired on television (antenna), video rushes compiled in a continuous sequence, a comprehensive image bank with traceable metadata from shooting, infographics and palettes, web modules, and internet subjects. Textual sources encompass press articles, PDF documents, internet links, and potentially contact information. For still images, the indexing process covers photographs, internet screenshots, title graphics (books, posters), as well as infographics, cartography, and iconography. The subjects to be indexed include those from the five national editions: 12/13h (France 3), 13h (France 2), 19/20h (France3), 20h (France 2), 23h (France 1), special programs, EVN, video rushes, as well as POOL and VO (voyage officiel).

Once the subjects are broadcasted, the INA (*Institut National de l'Audiovisuel*) retrieves them under a pre-existing convention between them. As per this agreement, after one year, the content is transferred to the INA and no longer belongs to France TV. Indexing of subjects and political interventions takes place one day after broadcast, while the indexing of pool and VO rushes is completed within one month by the respective services. The indexing of a TV news program takes approximately one day, with the 8 p.m. edition requiring more time, around 5 hours, compared to the 12:30 p.m. edition. These indexing procedures are crucial for organizing and cataloging the extensive content produced by France TV, facilitating efficient retrieval and future utilization of the archived materials.

2.3.3 Tools for content management and organization

France TV utilizes a diverse range of tools to effectively manage and organize content, adapting the tool selection based on specific regional requirements. In the Paris region, the primary tool employed for indexing is DALET. This advanced solution serves as an end-to-end, multiplatform news production and distribution tool, offering a comprehensive set of features tailored to the unique demands of the media industry. One of the notable advantages of DALET is its ability to seamlessly integrate new features, ensuring the smooth integration of our solutions and keeping pace with evolving needs.

In other regions of the metropole, a different indexing tool, called Sierra, is preferred. For indexing in the overseas departments and territories (DOM-TOM), Warehouse is the chosen tool, while in some overseas regions, Spring is still utilized, albeit considered outdated. During the elections, the documentalists use ELEC+. It provides a comprehensive database containing information about candidates, politicians, their biographies, photos, mandates, election scores, and results for both national and overseas elections. This tool serves as a valuable resource for documentalists to access relevant and up-to-date information about political figures and election-related data.

The indexing system's current structure comprises multiple local databases that are consolidated into a centralized national database. This setup facilitates comprehensive access to indexed content across different regions, ensuring efficient search and retrieval. By harnessing the capabilities of these tools and the integrated database, France TV streamlines content management and enhances organizational efficiency throughout its operations.

The content management process encompasses two crucial elements: indexing and search. While indexing focuses on organizing and categorizing content, it is essential for documentalists to keep the ultimate objective in mind: facilitating efficient discovery of the media during the search phase. To achieve this, documentalists adopt a user-centric approach and ask themselves, "If I were the one searching, what would I look for?". This perspective highlights the significance of understanding the main research categories and prioritizing the most frequent research queries. In the subsequent section, we will delve into these key research categories and provide a comprehensive explanation of the top recurring research topics.

2.3.4 Search process for efficient content discovery

This section primarily centers on the search process as deployed in the Paris region, featuring the DALET tool. The primary reason behind this focus is the Paris region's pivotal role as the central hub of France TV's operations. With the main headquarters and a wealth of media resources located in this region, the documentalists in Paris bear the responsibility of managing and organizing a substantial volume of content that circulates throughout the network.

The DALET tool provides documentalists with seamless access to the vast archives where all the content is stored. Within this tool, documentalists have the ability to conduct various types of research based on their specific needs. There are two main types of search options available. The first type, illustrated in Figure 2.3.a, revolves around keyword-based searches. The keywords are derived from a well-defined vocabulary, which is the INA thesaurus. It includes glossaries such as geographic locations, personalities, thematic keywords, image keywords, and speakers. The second type of search, shown in Figure 2.3.b, is based on free-form text, allowing documentalists to perform searches using criteria such as complete text, title, subtitle, and full description. This flexible search option enables documentalists to explore the archives using more open-ended and context-specific queries.

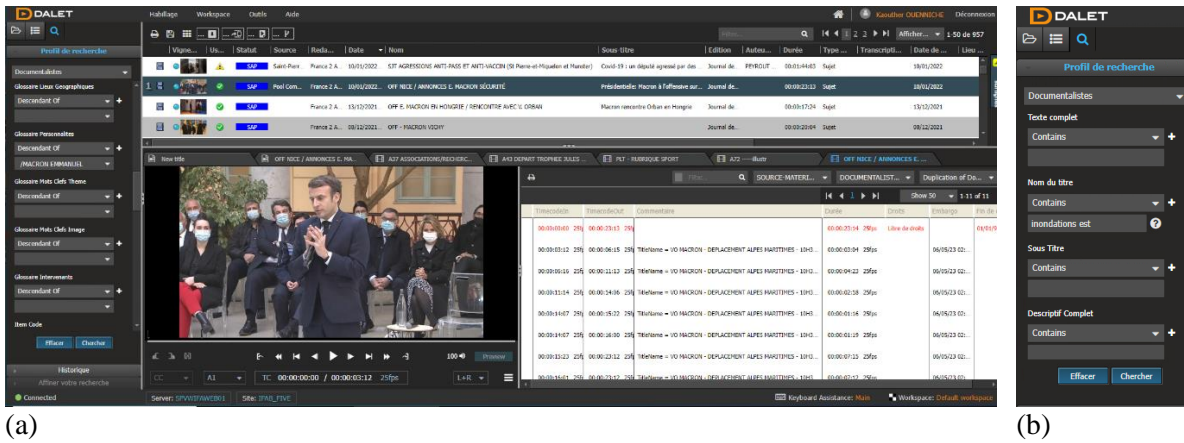


Figure 2.3. Overview of the DALET tool for content organization and management.

Documentalists conduct search queries to access pertinent content across various topics. The most frequent research themes include crowd scenes, traffic conditions, weather updates, transportation hubs, educational settings, smoking areas, shopping scenarios, fitness facilities, delivery services, governmental institutions, Paris landmarks, healthcare and COVID-related subjects, environmental concerns, political figures, web and technology matters, as well as leisure and cultural activities. In our research, we have conducted a survey among documentalists from the Club des Archives de l'Audiovisuel du Public.

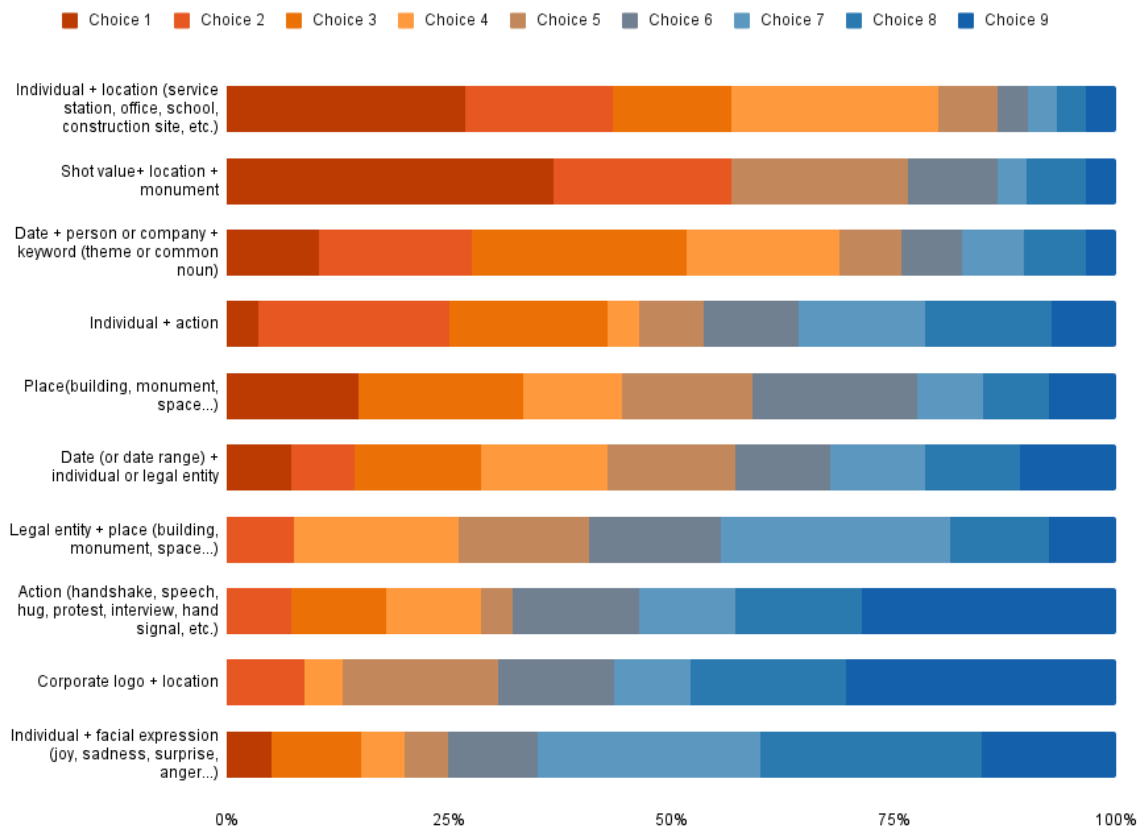


Figure 2.4. Top search items conducted by documentalists.

One of the key questions we asked them was to suggest the most frequent research items and prioritize them. The results of this survey are presented in Figure 2.4. Most of the research queries suggested by the documentalists involve a combination of multiple items such as Field of view shot type + place + landmark. An illustrative example of a query could be searching for "GP EXT PANO Notre Dame," which indicates a large shot capturing the exterior of Notre Dame Cathedral with a panoramic camera movement.

As part of our survey, we also sought to understand the most frequent actions that documentalists search for, as this is an integral part of their information needs. In Figure 2.5, we present the results of this inquiry. One notable action that emerged is "handshake", which exemplifies the type of action documentalists commonly seek. For instance, an illustrative query could be "Emmanuel Macron + shake + Vladimir Putin," indicating a search for footage capturing a handshake between Emmanuel Macron and Vladimir Putin. These insights shed light on the specific inquiries documentalists prioritize in their research efforts and inform the development of effective indexing and retrieval mechanisms for audio-visual content.

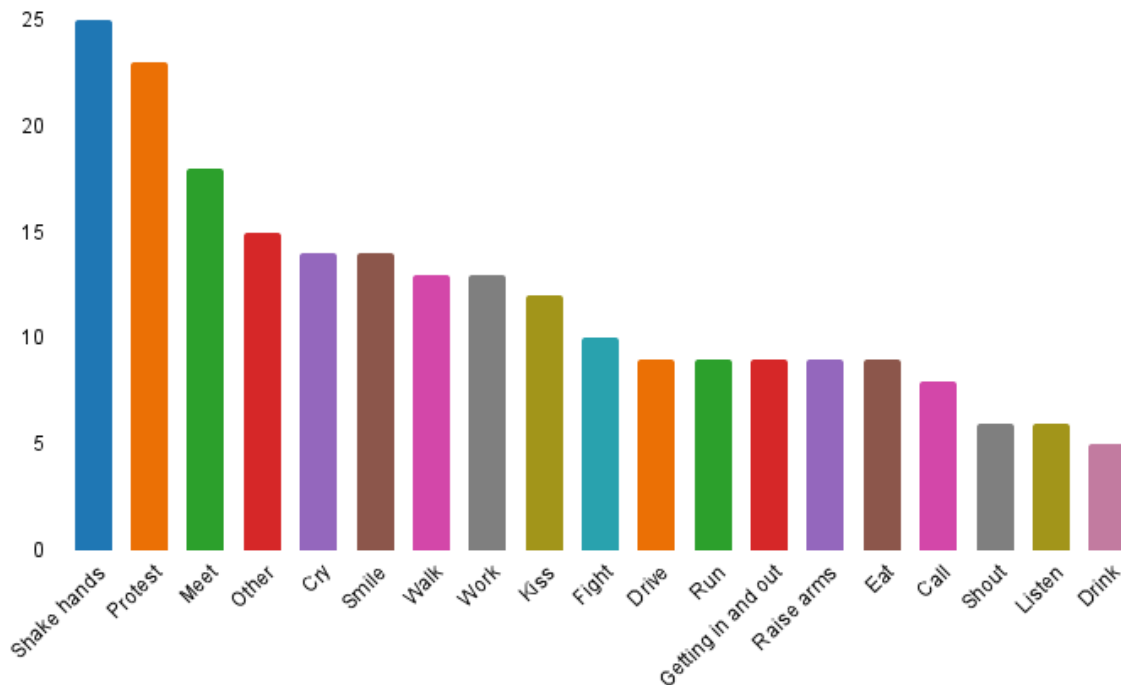


Figure 2.5. Top actions researched by documentalists.

The second element of content management is indexing. Documentalists play a crucial role in accurately indexing the subjects by utilizing a diverse range of metadata to capture and describe the intricate details of each shot and the overall subject. In the following section, we delve into the various types of metadata employed in this indexing process.

2.3.5 Overview of metadata for indexing

Metadata provides information on how, when, where, and by whom the data was collected, along with details about its availability, distribution method, projection system, coordinate system, tracking scale, resolution, precision, and reliability with respect to certain standards.

For example, when considering a news broadcast, metadata corresponds to the footage produced by journalists, which serves as the basis for creating news subjects. The main actors involved in the production of metadata are journalists and documentalists. Journalists handle the creation of news subjects, while documentalists qualify the metadata through indexing processes.

Each subject is indexed using a combination of keywords from the INA thesaurus. The thesaurus is organized into nine thematic "facets" corresponding to nine sections, where each term within them is hierarchically organized. For example, the keyword "aircraft carrier" is a "child" of "warship," which is a "child" of "naval equipment," and so on, ultimately leading to the broader category of "French Politics."

To classify each subject under a specific category, the thesaurus terms have been assimilated into a classification scheme, with each descriptor unequivocally associated with a particular category. This approach allows a correspondence table to be created between the thesaurus keywords and 14 predefined categories, ensuring that each subject is classified under the most relevant category. The INA's categories encompass a wide range of subjects, including Catastrophe, Culture and Leisure, Economy, Education, Environment, Crime and Accidents, History and Tribute, International, Justice, French Politics, Health, Science and Technology, Society, and Sports.

2.3.6 Types of metadata used for indexing

During the indexing process, a variety of metadata is used to categorize and describe each shot within a subject. In Figure 2.6, we show an example of an indexing page. The boundaries of each shot, indicated by the time code start (TC In) and time code end (TC Out), are pre-filled by the video editor and are specific to the source of the shot (e.g., France TV, INA archives, web). For each shot, documentalists define the following values:

- Field of view shot type (e.g., *Plan Moyen* (PM):medium shot, *Gros Plan* (GP):wide shot)
- Camera movement (e.g., zoom, traveling, panoramic),
- Location (interior (INT)/exterior (EXT)) and specific places (kitchen, café, conference room),
- Geographical location (e.g., Lyon, Paris, Bordeaux),
- Personalities (e.g., Emmanuel Macron, Elisabeth Borne) and speakers,
- Free-form text summary of the shot's content.

Liste séquences			
TC In TC Out Durée	Descriptif	Source	Droits
00:00:00:00 00:00:02:15 00:00:02:15	ARCHIVES PM Femme anonyme.	ARCH 10-07-2020 IV3 TOULOUSE S_ARCHITECTES_TRAVAIL_DISSIMULE, 2020-11-19 16:36:27, IV3_TOULOUSE_S_ARCHITECTES_TRAVAIL_DISSIMULE, Autre, IV3	FTV
00:00:02:16 00:00:09:14 00:00:06:23	ARCHIVES, 26/05/2020 ITW d'une personne anonyme à son domicile : "Je travaille à 100% et j'ai même dû travailler certains week-ends."	C12 - LES ABUS DU CHÔMAGE PARTIEL	FTV
00:00:09:15 00:00:13:10 00:00:03:20	GP Planning de télétravail sur un ordinateur.	Télétravail SXS Carte01 - 001, 2020-09-11 12:40:07, GLX_12247611, Rush	FTV
00:00:13:11 00:00:19:04 00:00:05:18	ARCHIVES ITW par téléphone d'une personne anonyme : "Les jours où j'ai travaillé, où j'ai fait des journées de 7 heures, ils m'avaient mis en chômage partiel."	ILLUSTRATIONS TEL 2020-09-17 17:38:48, GLX_12274310, RushBIGARD_J_Fraude_au_chômage_partiel_3, 2020-09-17 17:19:50	FTV
00:00:19:05 00:00:23:24 00:00:04:19	ARCHIVES avec effet, 16/03/2020 Allocution d'Emmanuel MACRON	16 03 2020 CORONAVIRUS: ALLOCUTION EMMANUEL MACRON [00:43:45:08 - 00:44:49:15] [extract], 2020-11-19 16:22:23, GLX_12538079, Compilation	
00:00:24:00 00:00:39:10 00:00:15:10	EXT DP Ministère du Travail.	IMAGES D'ILLUSTRATION : Ministère du Travail [00:00:00:00 - 00:04:42:17] [extract], 2020-11-19 15:40:08, GLX_12537754, Rush	FTV
00:00:39:10 00:01:04:24 00:00:25:14	INFOGRAPHIE Escroquerie - Usurpation d'identité d'une entreprise	INFOG_SCHLIENGER_ESCROC_CHOMAGE_PARTIEL_1, 2020-11-20 11:28:48, GLX_12541552	FTV

Figure 2.6. Example of an archive file. Source: DALET.

Furthermore, documentalists tag additional keywords to provide a quick way to retrieve content, as illustrated in Figure 2.7. The "*mots-clés images*" correspond to the thesaurus keywords that describe each shot, such as origin-diversity (presence of people of color in the shot), woman, laboratory, and more. Additionally, "*mots-clés thèmes*" are assigned to each subject, providing keywords that describe the overall content, such as medical research and health, also part of the thesaurus. If landmarks are present in the shots, they are noted as well. This is particularly valuable for events like the "*Tour de France*" which contributes to the promotion of tourism in France. These metadata elements ensure a detailed and comprehensive indexing of each subject, facilitating efficient retrieval and categorization of the content.

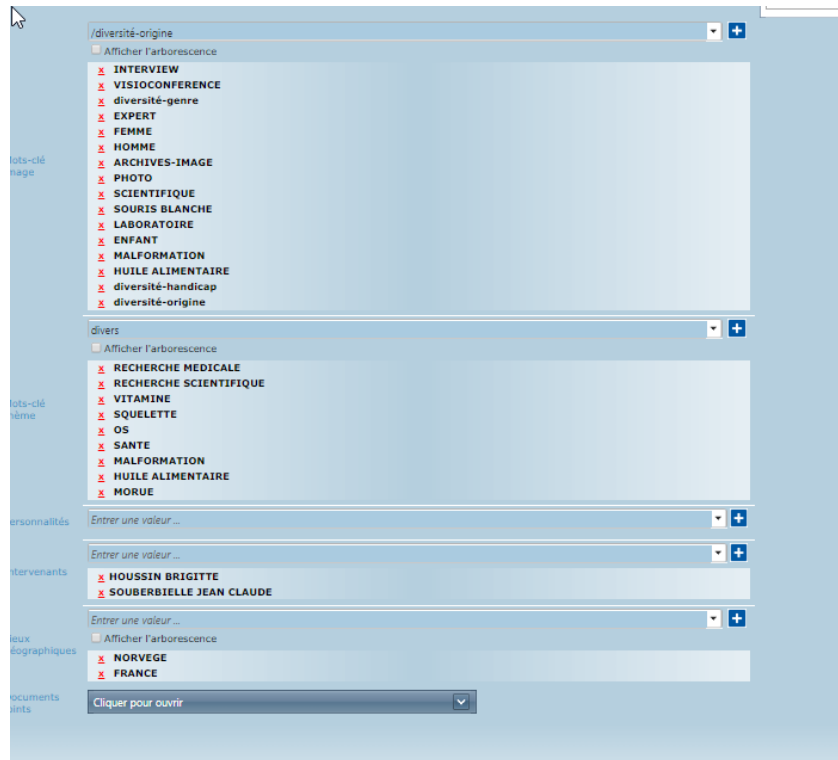


Figure 2.7. Example of tagged-keywords in an archive file. Source: DALET.

For official speeches by the president or other government representatives, it is important to provide the transcript of the speech as part of the indexing process. During elections, documentalists can also index additional metadata for each candidate, such as their political party affiliation, campaign promises, previous electoral performance, and public statements (see Figure 2.8). Furthermore, documentalists may be tasked with measuring the speaking time of each participant during debates, ensuring a fair representation and analysis of the discussions.

Figure 2.8. Index page for election candidates. Source: ELEC+.

2.4 Automation techniques

The task of indexing involves meticulously reviewing and categorizing vast amounts of content, identifying key elements, and assigning appropriate metadata. This manual process can be laborious, as it requires documentalists to carefully analyze each piece of media and make subjective decisions about the relevant tags and descriptors. Moreover, inconsistencies may arise among different documentalists, as their interpretations and perspectives may vary, despite the common INA thesaurus used. The resulting variations that appear during the indexing process can make the retrieval of specific content more difficult. Automating the indexing process in a comprehensive and objective manner can help standardize the indexing and ensure consistency. Additionally, the manual indexing process is susceptible to human errors and can be overwhelming, particularly during periods of high workload. Finding solutions to streamline and optimize this process is crucial for improving efficiency and enhancing the accessibility and retrieval of indexed content. We discuss in this section the possible solutions to streamline the documentalists' work.

2.4.1 Automatic solutions to assist the documentalists in archive management

Currently, for official speeches of government representatives, documentalists rely on searching for the streamed YouTube videos of these speeches and manually replicating the generated transcripts. However, this approach is not always available, as not all discourses are streamed or easily accessible. Furthermore, the manual replication process is prone to errors, which can lead to inaccurate indexing and hinder efficient retrieval of specific content. By leveraging speech-to-text technology, documentalists can automate the transcription process, converting spoken words into text with high accuracy. This not only saves time but also ensures the availability of reliable and searchable transcripts for indexing purposes.

Another challenge faced by documentalists is the difficulty of identifying the original source of videos in the archives, especially when the metadata indicating the source is lost or incomplete. This poses a problem, particularly for media that may have been purged or removed from official channels. To address this issue and ensure compliance with copyright regulations, fingerprinting technology can be employed. By applying fingerprinting algorithms, documentalists can compare the visual characteristics of videos in the archives with known copyrighted content. This helps identify images that France TV does not have the right to use, preventing any potential copyright infringement.

The utilization of Optical Character Recognition (OCR) techniques brings numerous advantages to the indexing and retrieval process. Documentalists can leverage OCR technology to extract text information from various sources, such as the publicity or news subject displayed in the lower corner of a video. By automatically capturing this text, documentalists can enhance the indexing process by including specific keywords and metadata related to the content. Additionally, OCR proves invaluable in digitizing press reviews and old archives that exist in scanned formats. The OCR algorithms can analyze the scanned documents, recognize the text within them, and convert it into editable and searchable digital text. This enables documentalists to perform comprehensive searches directly within the digitized press reviews, eliminating the need for manual classification or physical archiving.

Thumbnail generation and shot boundary detection play a crucial role in the indexing and retrieval of video content, offering significant benefits to documentalists. Automating these processes can effectively address the challenges faced by documentalists who currently need to manually correct time codes provided by video editors. Thumbnail generation involves creating representative images that visually summarize the content of a video segment. The thumbnails serve as visual cues that aid in quickly identifying and navigating through videos, saving documentalists time and effort. Moreover, shot boundary detection automates the identification of scene transitions within a video, accurately marking the boundaries between shots. By automatically detecting shot boundaries, documentalists can rely on precise time codes, reducing errors and ensuring accurate indexing.

Archive file indexing involves the categorization and tagging of various elements within the videos to enable efficient retrieval and analysis. Tags such as camera movement, places, and field of view shot type play a vital role in indexing the media subjects. These tags should be identified shot by shot. However, manually indexing such elements can be time-consuming and resource-intensive. Deep-learning techniques have emerged as a valuable solution to streamline this process, significantly reducing the time and effort required. By leveraging these algorithms, archival file indexing can be expedited, allowing for faster and more accurate retrieval of specific shots and relevant information.

Furthermore, the documentalists are tasked with providing a free-form natural-language text that summarizes the content of each shot. This process is time-consuming, resource-intensive, and often subjective, leading to inconsistencies and potential errors in the indexing process. However, the emergence of AI solutions, specifically automatic video captioning, presents a transformative

opportunity. By harnessing the power of advanced algorithms in computer vision and natural language processing, these systems can automatically generate accurate and detailed captions for each shot, alleviating the burden on documentalists. This not only streamlines the indexing process but also ensures a more standardized and reliable approach, enhancing the overall quality and efficiency of video content management.

The *Tour de France* holds immense importance as it serves as a prominent platform for showcasing the diverse attractions, cultural heritage, and scenic landscapes of France. During this iconic cycling race, documentalists play a crucial role in capturing and indexing the footage. Currently, documentalists rely on a book that provides information about the landmarks in each region along the race route. Additionally, they have access to the dates on which the cyclists pass through these regions. The manual task of identifying and labeling shots based on the corresponding landmarks and dates in each region can be incredibly time-consuming, often requiring extensive research and meticulous attention to detail. However, leveraging automation and incorporating relevant landmarks from France can significantly alleviate this burden.

Documentalists can also use video question answering techniques to enhance the indexing process by automatically assigning relevant tags, categories, and keywords to videos based on the extracted answers. This streamlines the indexing workflow and ensures that videos are accurately labeled and classified for easier retrieval. Moreover, video question answering techniques enable documentalists to perform more precise searches. Instead of relying solely on manual annotations or limited metadata, they can now input specific queries and receive relevant video segments as results. This saves time and effort by directly retrieving the desired content, enhancing the overall efficiency of the indexing and retrieval process. Additionally, video question answering techniques can assist documentalists in uncovering hidden or hard-to-reach information within videos. They can ask complex questions about specific events, objects, or actions occurring in the video, and the system will analyze the content to provide detailed answers. This allows documentalists to discover valuable insights and relevant moments in the video, enriching the indexing process with deeper context and information.

2.4.2 Proposed solutions

Our PhD research was conducted in close collaboration with the DAIA (Data and AI) department at France TV. As part of our efforts to develop a comprehensive framework for multimodal indexing of television audio-visual content, various solutions were proposed from different teams within the group. These solutions included implementing automatic speech-to-text, OCR detection, thumbnail generation, and shot boundary detection.

Within this larger context, our specific focus was put on the indexing of the archival page, illustrated in Figure 2.6. We specifically aimed to align our work with the frequently expressed needs of documentalists (Figure 2.4). By understanding the challenges and priorities identified by documentalists in their day-to-day activities, we were able to tailor our efforts to address these needs effectively.

To achieve this, we have first proposed a set of automatic techniques to identify relevant keywords such as landmarks, camera movement, field of view, and place recognition. In the subsequent phase, we integrated multiple analyzers, leveraging video question answering technique to enhance scene understanding, which is considered an AI-complete task. Lastly, we employed video captioning techniques to provide a natural language description of the semantic content for each shot, further enhancing the indexing process.

2.5 Conclusion

In this chapter, we have provided a comprehensive overview of the context and objectives of our doctoral project. We have delved into the critical role that documentalists play in archive indexing and retrieval at France TV, offering an insightful look at their intricate and demanding responsibilities.

Throughout our discussion, we have explored the manifold challenges that documentalists routinely face in their work. A crucial aspect of their tasks, the process of manual indexing and retrieval, proves to be time-consuming and labor-intensive. Moreover, the work is not immune to the pitfalls of subjectivity and human bias, often resulting in inconsistencies in the data. Additionally, they operate under the constraints of limited time and resources, further compounding the difficulty of their duties. To shed light on these issues, we conducted a comprehensive study comprising multiple interviews with a diverse range of stakeholders in the broadcast community at France TV and its associated partners. These interactions served to identify and evaluate potential solutions to mitigate the aforementioned challenges.

In the next chapters, we detail the frameworks we have considered for each task. We have chosen to index the following metadata:

- landmarks, which signify the cultural heritage of France;
- camera motion and field of view shot types, which serve to classify shots according to the aesthetic and artistic choices made by the videographer;
- video captioning, to generate a concise, free-form natural language text that encapsulates the content of each shot;
- and finally, Video Question Answering to streamline the retrieval process without necessitating an exhaustive list of metadata during the indexing process

We categorize the tasks based on the modalities used. Unimodal models, discussed in the first part of the manuscript, use a single modality like image, text, or video, and include the following tasks: landmark recognition, scene segmentation, and camera motion categorization. Conversely, multimodal models, covered in the second part, merge various modalities to gather information and are involved in the video question answering and video captioning tasks.

PART I: UNIMODAL, TASK-SPECIFIC MODELS

3 LANDMARK RECOGNITION

Abstract: Landmark recognition and retrieval plays a pivotal role in the indexing of television archives, offering an effective mechanism to categorize and access vast volumes of visual data. In this chapter, we detail our work for landmark recognition, with specific application to the needs of France TV. We employ a recognition-by-retrieval approach to contend with several challenges, including dataset variance. Further, we introduce an automatic technique for the construction of a landmark database for France TV. This technique utilizes the organization's thesaurus, ensuring the database is well-tailored to its specific archival requirements. We test our approach in zero-shot setting on well-known benchmarks and achieve competitive results.

Keywords: Landmark recognition, content-based retrieval, archive indexing.

3.1 Introduction

Major events like “Tour de France” hold immense importance as they serve as prominent platforms for showcasing the diverse attractions, cultural heritage, and scenic landscapes of France. Currently, documentalists heavily rely on a reference book that provides information about the landmarks situated in each region through which the cyclists pass. This information, coupled with the dates of the race in these regions, assists documentalists in identifying and labeling shots based on the corresponding landmarks and dates.

To streamline and automate this laborious task, we propose leveraging deep learning-based automation techniques for landmark recognition. By employing advanced computer vision algorithms, we aim to develop a system that can automatically identify landmarks and associate them with the corresponding regions and dates. The foundation of our landmark recognition system lies in the geographical thesaurus of INA, which encompasses 139 landmarks situated in different cities around the world. This comprehensive list of landmarks will form the core of our landmark dataset, as explained in detail in section 3.3.

The remainder of the chapter is structured as follows. First, we review the latest developments in the area of landmark recognition and retrieval, emphasizing the related challenges and limitations. We then classify the existing methods into two groups: traditional strategies and deep learning-based techniques. Next, we present our approach to automatically generate a dataset specifically tailored for France TV applications. In section 3.4, we explain our approach for landmark retrieval delving into the network architecture, the learning procedures, and the processes used for image retrieval. Finally, we show the effectiveness of our method through tests and results from our dataset, as well as known benchmarks in a zero-shot setting. We compare our approach to previous state of the art methods. Finally, we conclude the chapter and present some possible future projects.

3.2 Related work

Landmark recognition in computer vision poses several challenges due to the inherent complexity and variability of landmarks. In this section, we explore the existing literature on landmark recognition, highlighting the difficulties faced in this domain.

3.2.1 Challenges in landmark recognition

Landmarks encompass a wide range of structures, including historical sites, natural landmarks, modern architectural marvels, and cultural artifacts. Unlike well-defined objects, landmarks do not have a formalized concept or a specific set of visual characteristics that universally define them. Instead, they exhibit significant variations across civilizations, regions, and historical periods. This diversity poses a considerable challenge in developing a generalized approach for landmark recognition. For instance, the architectural styles and cultural significance of landmarks can vary significantly between different countries or even within the same country. This variability requires landmark recognition systems to be adaptable and capable of capturing the unique visual characteristics of each landmark type.



Figure 3.1. Variations in viewpoint, illumination and presence of distractors.

Landmark recognition is further complicated by variations in viewpoint, illumination conditions, image resolution, and the presence of distractors (Figure 3.1). When capturing landmarks from different angles or distances, the appearance and visual cues may change dramatically. Illumination conditions can introduce shadows, highlights, or occlusions, altering the appearance of the landmarks. Moreover, images obtained under adverse weather conditions or low-light environments can affect the visibility and quality of the captured landmarks. Additionally, the presence of distractors such as trees, people, or other structures in the vicinity of landmarks can impede the learning process. Distinguishing the landmark of interest from its surroundings becomes more challenging when such distractors partially occlude the landmark.

On the counterpart, certain types of landmarks, particularly churches and other architectural structures, can exhibit great resemblance to one another. This leads to low inter-class variability, making it difficult for recognition algorithms to differentiate between similar landmarks. This problem becomes even more pronounced when landmarks share common architectural styles or have similar visual features. Figure 3.2 illustrates the challenges of low inter-class variability. In such cases, the discrimination between similar landmarks requires the utilization of subtle visual cues, contextual information, and a deep understanding of architectural nuances.

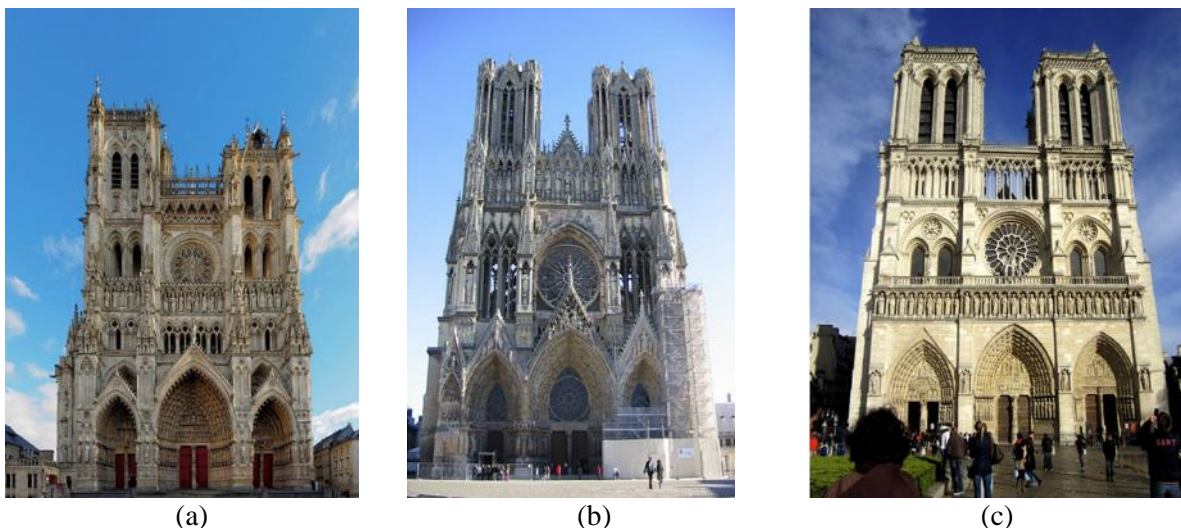


Figure 3.2. Three look-alike gothic churches. (a) Notre Dame, (b). Amiens, (c). Reims.

In [5], the authors propose a recognition-by-retrieval technique, wherein the focus is shifted from directly predicting the class of a query image to applying a majority-voting scheme on the top-five

nearest extracted images. Inspired by this approach, we adopt a similar strategy in our work and treat landmark recognition as an instance retrieval task.

3.2.2 Content-based image retrieval

Addressing the problem of landmark recognition as an image retrieval task is beneficial for different reasons. One of the key benefits is the flexibility, since it becomes possible to add/remove classes without the need any retraining scheme. This allows for efficient and dynamic management of the landmark classes, accommodating changes and expansions in the thesaurus landmark database without significant computational overhead. Furthermore, applying a majority voting scheme on the top-k nearest neighbors in the recognition-by-retrieval approach helps to mitigate potential errors. By considering multiple nearest neighbors and their associated probabilities, we can account for situations where the probabilities of the first few images are close to each other (Figure 3.2). This approach helps us reduce the impact of misclassifications or uncertainties in the prediction process, thereby enhancing the robustness and reliability of the landmark recognition system. The incorporation of image retrieval techniques in this context allows us to leverage the collective knowledge of the nearest neighbors. An image retrieval framework involves the use of a query image to search for relevant or similar images within a database (Figure 3.3). Given a query image, the framework first processes this image through a set of procedures. A first stage involves the feature extraction process, used to describe/characterize the image. The features can include color and texture descriptors (section 0) or patterns directly learnt by convolutional neural networks (3.2.2.2), in the case of deep learning approaches. The features are compared with the help of a dedicated similarity measure, which most of the time is defined in terms of distances in the feature space. The retrieved images are finally ranked according to their similarity to the query.

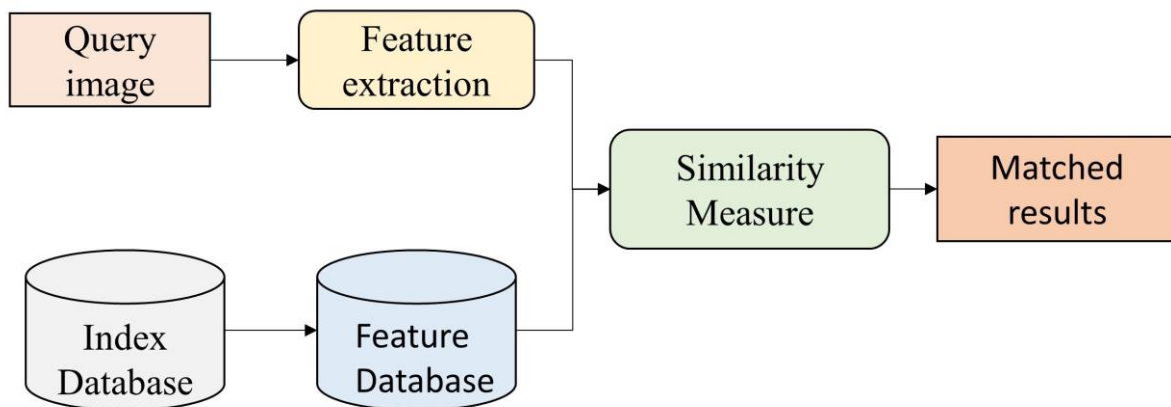


Figure 3.3. Overview of the image retrieval framework.

3.2.2.1 Traditional methods

Content-based image retrieval has been an active research area for decades, with early works focusing on extracting visual cues such as color [6], shape [7], texture [8], and spatial information [15]. Various algorithms have been proposed to address these aspects and improve retrieval performance. In his seminal work, Lowe introduced the so-called scale invariant feature transformation (SIFT) descriptor [9], which capture the local texture information around a set of interest points. By providing

local representations that are robust to changes in scale, rotation, and viewpoint, the SIFT approach paved the way for subsequent advancements in the field of image retrieval.

Traditionally, image retrieval methods begin by extracting reliable interest points that are invariant to large viewpoint changes using local detectors such as the difference of Gaussians (DoG) [10], Harris-Laplace [11], or Hessian-affine [12]. These interest points serve as key locations within the image that encapsulate distinctive visual information. The content of the local regions surrounding these interest points is then encoded using feature descriptors.

In addition to SIFT, other notable descriptors have been proposed to enhance efficiency and performance. PCA-SIFT [13] combines principal component analysis with SIFT to reduce the dimensionality of the feature vector, improving computational efficiency while maintaining accuracy. RootSIFT [14] normalizes the SIFT descriptor to make it more robust to varying illumination conditions. SURF (Speeded-Up Robust Features) [15], an alternative to SIFT, integrates the Hessian-Laplacian detector to achieve comparable accuracy with reduced computational time.

To aggregate the vector-based descriptors, various techniques have been introduced. Bag of words (BoW) [16] models the distribution of visual words in an image, treating it as a histogram-like representation. Fisher vectors (FV) [17] encode higher-order statistics to capture more fine-grained information. Vector of locally aggregated descriptors (VLAD) [18] aggregates the differences between descriptors and cluster centers, providing a more discriminative representation.

It is worth noting that these traditional methods have shown effectiveness in content-based image retrieval. However, they often suffer from limitations in dealing with large-scale datasets, high-dimensional feature spaces, and complex semantic understanding. Hence, the recent advancements in deep learning-based approaches have gained significant attention in addressing these challenges and pushing the boundaries of landmark recognition and image retrieval tasks.

3.2.2.2 Deep learning-based techniques

Since the introduction of AlexNet [19] in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012, convolutional neural network (CNN) models have surpassed hand-crafted features, demonstrating their effectiveness in extracting robust features from raw images. In [20], [21], the activations of fully connected layers are used as global descriptors, achieving promising performance under an Euclidean distance, which can be further improved using power normalization [17]. Moreover, in [22], [23], authors directly extract the output of intermediate convolutional layers as descriptors of image patches corresponding to receptive fields of the features. These convolutional kernels serve as local descriptors that are more resilient to occlusion, truncation, and clutter. While these descriptors can be used off-the-shelf, better results can be achieved through network initialization, domain adaptation, and transfer learning. The standard architecture for instance retrieval is the Siamese network, also known as a twin network, which learns semantic similarities between matching and non-matching pairs for model training. It has been demonstrated [24] that twin networks are more robust to class imbalance, a common characteristic of landmark datasets. Siamese networks consist of two or more branches, where each branch shares the same configuration and weights. The goal is to learn the similarities between a pair of images by comparing their corresponding feature vectors. This is accomplished by feeding the network a pair of images (i, j) and their corresponding class labels $Y(i, j) \in \{0, 1\}$, indicating whether the pair is non-matching (label 0) or matching (label 1), and training the model using a contrastive loss (Figure 3.4).

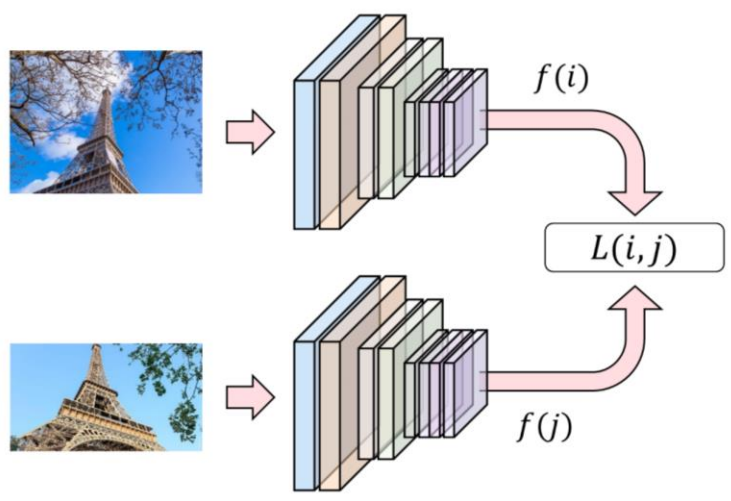


Figure 3.4. Siamese network architecture with contrastive loss.

Another approach exploits the so-called the triplet loss, which utilizes a triplet of images to compute similarity rankings. Instead of comparing similar and different images, the triplet loss minimizes the distance between an anchor image (A) and a positive image (P) from the same class while maximizing the distance with a negative image (N):

$$L(A, P, N) = \max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0) \quad (3.1)$$

where f denotes the function that associates a feature to a given image and α is a margin between positive and negative pairs.

The set of local features extracted from the convolutional layers is then aggregated into a global descriptor. Experimental results in [21], [25], [26] have shown that simple aggregation methods such as max pooling or sum pooling outperform traditional techniques like Vector of Locally Aggregated Descriptors (VLAD) or Fisher Vectors (FV).

Our work draws inspiration from the success of Convolutional Neural Networks, specifically implementing the ResNet [27] model as a backbone architecture. We employ a training method that utilizes contrastive loss within a siamese network. To reduce the computational overhead, we apply Principal Component Analysis (PCA) to reduce the dimensions of the resultant feature vectors. Subsequently, we identify the most similar images to the query utilizing the K-Nearest Neighbors algorithm. During the evaluation of our model, we confronted challenges concerning existing benchmarks, primarily in relation to their scale and generalizability. To overcome such limitations, we have automatically constructed a dataset for validation, as described in the following section.

3.3 Constitution of a landmark dataset

Creating a suitable dataset is crucial for training and evaluating deep learning models for landmark recognition. While there are commonly used benchmarks like Paris6k [3], Oxford5k [28], and INRIA holidays [29], these datasets have limitations in terms of size and generalizability. On the other hand, the Google landmark dataset [30], with over 5 million images and 200,000 instances, is large but presents challenges such as noise, imbalance, and high computational requirements. Moreover, its coverage extends to cities not included in the geographic thesaurus of INA, making it less relevant for the indexing process of France TV archives. Therefore, to address these issues, we have generated a landmark dataset tailored to the specific requirements of documentalists.

Creating a custom dataset poses the challenge of avoiding bias in image selection and labeling. Subjective criteria such as specific viewpoints or occlusion should be minimized. To achieve this, we propose utilizing automatic techniques that leverage the metadata and geo-coordinate locations available from the Flickr API . [31]Our approach involves the following steps. First, we compile a list of cities and their corresponding landmarks from INA's geographic thesaurus. Next, we search for photos on Flickr that include the landmark name and city as tags, aiming to avoid confusion with landmarks of the same name in different cities. To filter out potentially inaccurate or irrelevant photos, we dismiss those with geotag precision scores below a threshold (13 in Flickr metadata), which corresponds to a precision superior to the one of a city block. However, it is important to note that these keywords available on Flickr images are user-generated and do not guarantee the presence of the landmark in the photo, which can introduce noise into the resulting database. Figure 3.5 illustrates some examples of falsely tagged landmarks.

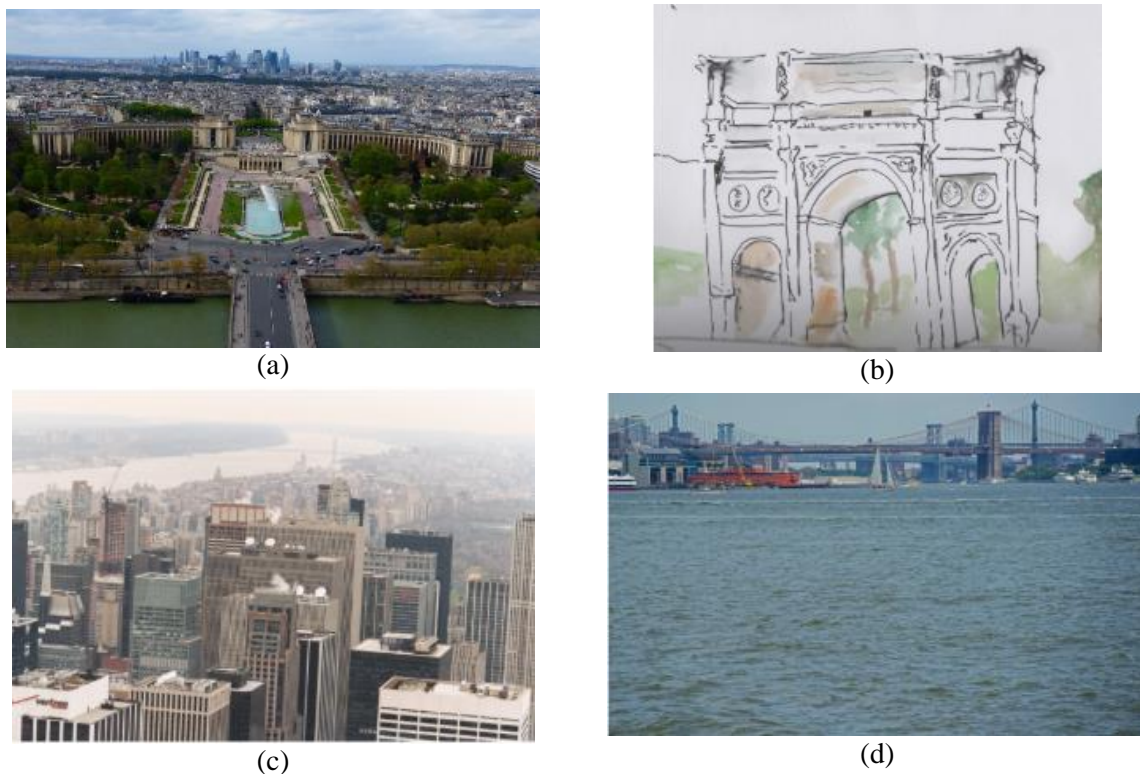


Figure 3.5. Examples from users' pictures falsely tagged as a landmark. (a) Eiffel tower (b) Arch of Constantine (c) Empire state building (d) Statue of liberty

To mitigate this problem, we begin by extracting the latitude and longitude coordinates, representing each landmark as a point on a two-dimensional plane. Subsequently, we employ mean-shift clustering [32] to discern the centroid of each cluster. By leveraging this technique, we effectively identify the central point around which the landmark's images are concentrated. Next, we retrieve images from Flickr, focusing on a radius of 500 meters around each centroid. These images are obtained in Flickr's medium resolution, ensuring a balance between image quality and computational efficiency. The resulting dataset encompasses a total of 139 landmarks, with an average of 200 images per landmark class. Figure 3.6 presents some examples from the dataset. In Table 1, we present the geo-coordinates of the centroids for five selected landmarks.



Sagrada Familia



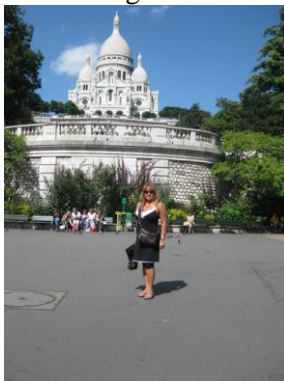
Empire State building



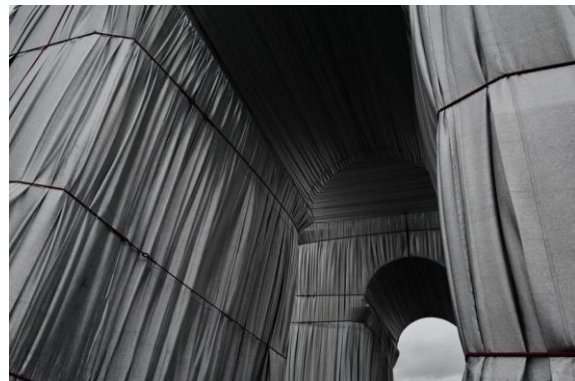
Big Ben



Arch of Constantine



Sacre coeur



Triomphe



eiffel



Colosseum

Figure 3.6. Examples from the dataset.

Table 3.1. Geo-coordinates of five landmarks.

Landmark	Geo-coordinates	Tags
Eiffel tower	48.8583, 2.2942	'eiffeltower', 'paris'
Sacre coeur	48.8863, 2.3430	'sacrecoeur', 'paris'
Big Ben	51.5008, -0.1243	'bigben', 'london'
Sagrada familia	41.4036, 2.1742	'sagradafamilia', 'barcelona'
Colosseum	41.8904, 12.4920	'colosseum', 'rome'

3.4 Proposed methodology

Our approach consists of a three-step pipeline. Firstly, we present the network architecture for feature extraction and feature aggregation. Secondly, we train our model using contrastive loss with a siamese network. Lastly, we perform image recognition through a retrieval process, which will be detailed in section 3.4.4.

3.4.1 Network architecture and image representation

We have considered ResNet-101 [27] network pre-trained on ImageNet [19] as the backbone, enabling the generation of high-level feature representations. The architecture employs 2D convolutions and pooling operations. The CNN consists of 101 layers grouped into five blocks: conv1, conv2_x, conv3_x, conv4_x, and conv5_x. Each block involves successive convolution operations repeated $x = 3, 4, 2, 3,$ and $3,$ respectively. In our framework, we remove the fully connected layer at the end. The convolutional kernels have a size of 3×3 except for conv1 with a kernel size of 7×7 . Down sampling of inputs is accomplished by the conv2_1, conv3_1, conv4_1, and conv5_1 layers, employing a stride of 2. Following each convolutional layer, batch normalization and ReLU activation are applied. The output of the residual block is obtained by adding the input to the block with the output of the last convolutional layer of the block, forming a skip connection. The 2D-CNN receives an image pair as input, processes them, and extracts visual descriptors from the last convolutional layer (conv5_3). In Figure 3.7, we present an overview of the network architecture. Let $x_i \in \mathbb{R}^{224 \times 224 \times 3}$ be the input image, we denote by $Z_i \in \mathbb{R}^{H \times W \times C}$ the feature map extracted from the conv5 layer where $H \times W = 7 \times 7$ represents the spatial dimensions, and $C = 2048$ denotes the number of channels capturing various feature representations.

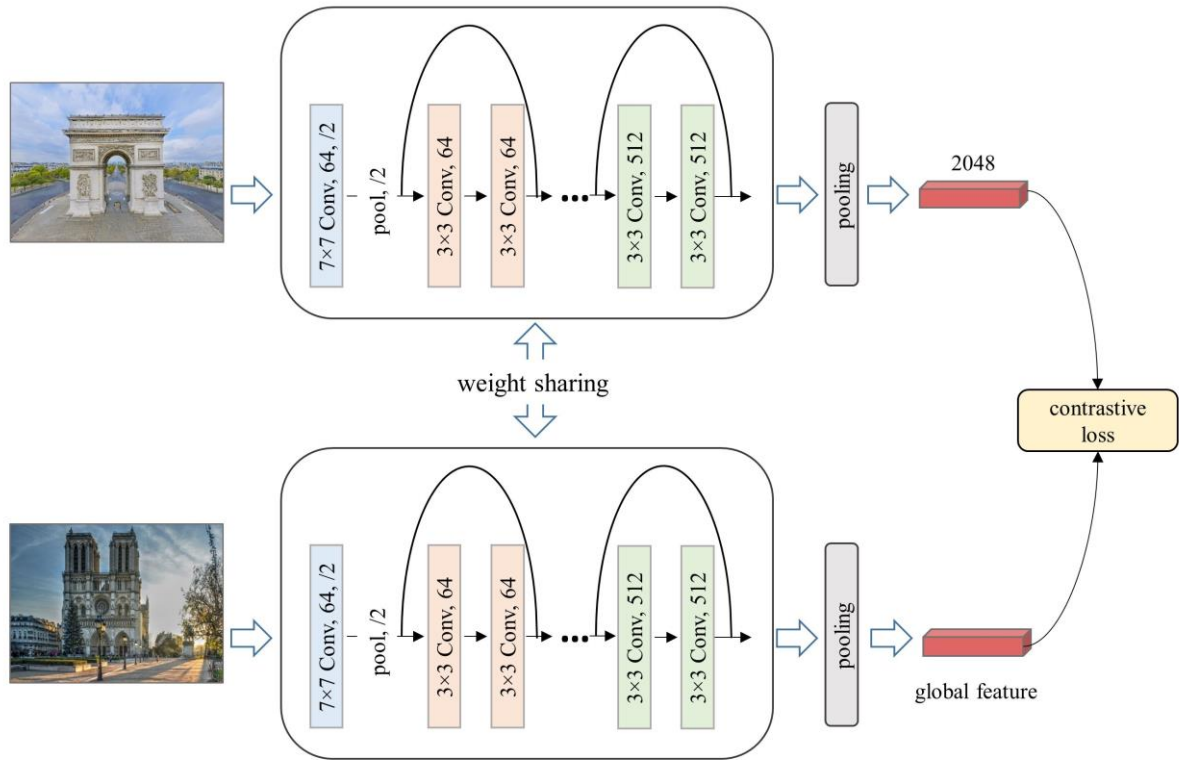


Figure 3.7. Network training using contrastive loss (offline).

Let A_c be the set of all $H \times W$ activations for feature map $c \in \{1, \dots, C\}$. To aggregate the features, we use the Maximum Activations of Convolutions (MAC) [24] technique (Figure 3.8). The global feature vector is constructed by max-pooling over all dimensions per feature map. It is computed as:

$$\mathbf{Z}_i = [Z_{1,i} \dots Z_{c,i} \dots Z_{C,i}]^T, \text{ with } Z_{c,i} = \max_{a \in A_c} a \cdot \mathbb{1}(a > 0) \quad (3.2)$$

where the indicator function $\mathbb{1}(a > 0)$ equals 1 when the value of a is greater than zero, and 0 otherwise.

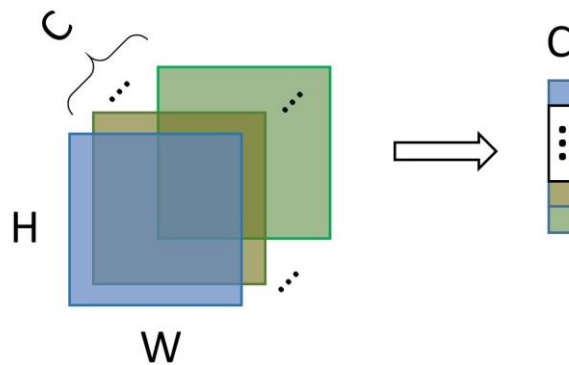


Figure 3.8. Feature aggregation using MAC technique.

3.4.2 Siamese learning

We address the problem of landmark retrieval as similarity learning task. The main objective of the task is to ensure that positive samples are positioned closer together in the latent space according to a defined metric, while negative samples are positioned farther apart (see Figure 3.9). The goal is to maximize the distinguishability between positive and negative samples, enabling the model to learn meaningful representations that accurately capture the similarity and dissimilarity relationships within the data.

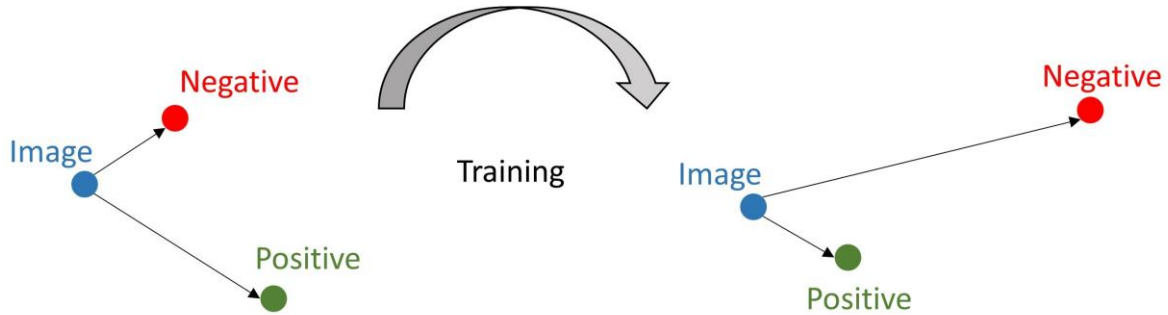


Figure 3.9. Similarity learning task. The objective is to minimize the distance between positive samples and to maximize the distance between negative samples.

For this purpose, we use Siamese network as illustrated in Figure 3.7. This network consists of two identical subnetworks, both utilizing the ResNet-101 architecture. The subnetworks extract features from the images x_i and x_j and map them into a high-dimensional feature space. A distance function is then employed to compute the similarity between the extracted features (\mathbf{Z}_i and \mathbf{Z}_j). Notably, the subnetworks share weights, ensuring that the two images are represented similarly in the feature space. During the training phase, we employ contrastive loss to optimize the model. By utilizing contrastive loss, we aim to bring similar images closer to each other in the feature space while pushing non-similar images further apart. The loss function takes as input the Euclidean distance between the features extracted from the two sub-networks. When the label of the two input images is 1 (indicating similarity), the loss function minimizes the Euclidean distance. Conversely, when the label is 0 (indicating non-similarity), the loss function encourages an increase in the Euclidean distance between the features. This approach allows the network to effectively learn to distinguish between similar and non-similar images based on their feature representations. The contrastive loss is defined as follows:

$$\mathcal{L}(\mathbf{Z}_i, \mathbf{Z}_j, y_{ij}) = (1 - y_{ij}) * D(\mathbf{Z}_i, \mathbf{Z}_j)^2 + y_{ij} * \max(\alpha - D(\mathbf{Z}_i, \mathbf{Z}_j), 0)^2 \quad (3.3)$$

where $y_{ij} = 1$ if the images are similar (positive samples) and $y_{ij} = 0$ if the images are dissimilar (negative samples); $D(\mathbf{Z}_i, \mathbf{Z}_j)$ represents the Euclidean distance between the feature vectors \mathbf{Z}_i and \mathbf{Z}_j ; and α is a hyperparameter that controls the desired separation between similar and dissimilar image pairs.

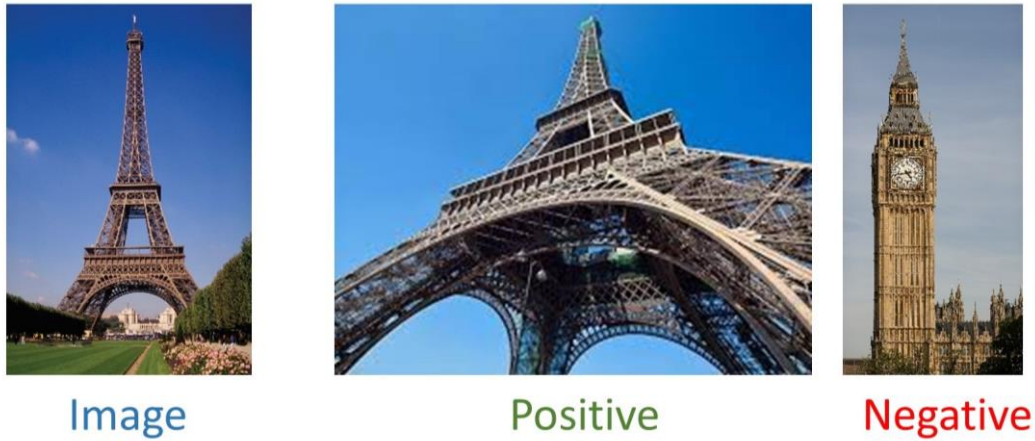


Figure 3.10. Example of batch-wise positive/negative mining.

To maximize the utilization of the dataset, a batch-wise positive/negative mining process is employed to construct all possible positive and negative training pairs within each batch. This process ensures that an equal number of images from each category are included in the batch. For each image in the batch, a miner is employed to construct training pairs. The miner randomly selects either an image from the same category as the positive sample or an image from a different category as the negative sample. This random selection process is repeated for all images in the batch, allowing the mining process to create all positive and negative pairs necessary for training. In Figure 3.10, we show an example of the mining process.

3.4.3 Dimensionality reduction and whitening

In this section, we present the post-processing of the global feature vector. The output of the CNN is a high-dimensional feature vector (2048 in this case) which is computationally expensive and may lead to the curse of dimensionality. For this purpose, we apply PCA to reduce the dimensionality while retaining the most important information. Besides, PCA helps in removing noisy or less informative information, allowing the retrieval process to focus on more discriminative features. We also apply PCA whitening to transform the feature vectors to have uncorrelated components with unit variance. This process can be beneficial in image retrieval as it removes any linear dependencies between features, enhancing the ability of subsequent algorithms (k-NN in this case) to measure similarity accurately.

Formally, we normalize the feature vector \mathbf{Z} using ℓ_2 normalization. Our proposed approach is based on the work of [33]. The projection consists of two components: whitening and rotation. For the whitening part, we calculate the inverse square root of the intra-class (matching pairs) covariance matrix $C_S^{-1/2}$, with

$$C_S = \sum_{y_{ij}=1} (\mathbf{z}_i - \mathbf{z}_j)(\mathbf{z}_i - \mathbf{z}_j)^T \quad (3.4)$$

The rotation part involves performing PCA on the interclass (non-matching) covariance matrix in the whitened space, denoted as $eig(C_S^{-1/2} C_D C_S^{-1/2})$, where C_D is computed as follows:

$$C_D = \sum_{y_{ij}=0} (\mathbf{z}_i - \mathbf{z}_j)(\mathbf{z}_i - \mathbf{z}_j)^T \quad (3.5)$$

Then we apply the projection matrix $P = C_s^{-1/2} \text{eig}(C_s^{-1/2} C_D C_s^{-1/2})$ as $P^T(\mathbf{Z}_i - \mu)$, where μ represents the mean vector to perform centering. To achieve dimensionality reduction of the descriptors to d dimensions, we select only the eigenvectors associated with the d largest eigenvalues. The projected vectors obtained from this selection are then further processed by applying ℓ_2 -normalization. The resulting feature vector is denoted by $\tilde{\mathbf{Z}}$.

3.4.4 Image retrieval

Given an image query x_{query} , the objective of the system is to classify the label of the image. Leveraging the recognition-by-retrieval technique, the image retrieval phase involves identifying the k -nearest images to the query in the index database and determining the class of the query through a majority voting scheme. To accomplish this, we begin by extracting the feature vector of the query, denoted as \mathbf{Z}_{query} . This vector is obtained by applying ℓ_2 normalization to the MAC global feature extracted from the last convolutional layer of the CNN network. Subsequently, we perform dimensionality reduction using PCA, as discussed in Section 3.4.3. The resulting reduced-dimensional vector is denoted as $\tilde{\mathbf{Z}}_{query}$.

To identify the k -nearest images to the query, we utilize the k NN algorithm. The distance metric employed (Euclidean distance) quantifies the similarity between feature vectors. Given the reduced-dimensional feature vectors of the images in the index database, we compute the distances between the query and the images in the database. The k images with the smallest distances to the query are considered the k -nearest neighbors. Once the k -nearest images have been identified, a majority voting scheme is applied to determine the class of the query. Each neighbor contributes to the vote of its corresponding class label. The class with the highest number of votes is assigned as the predicted class for the query. Figure 3.11 illustrates the image retrieval process.

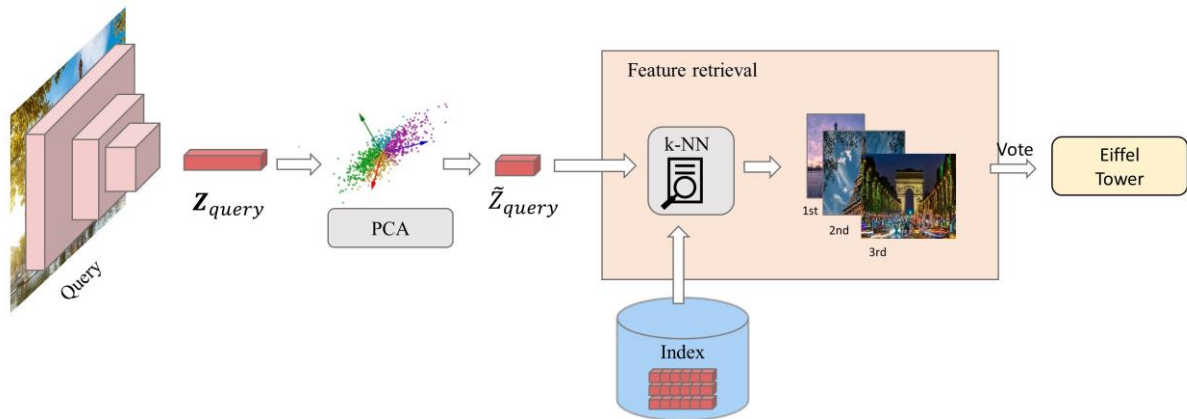


Figure 3.11. Overview of the image retrieval process.

3.5 Experiments and results

In this section, we present the datasets we used for training and validation. Next, we provide implementation details of our framework. We highlight the model results on the selected datasets along with a comparison with state-of-the-art methods. Finally, we present some qualitative results to further illustrate our findings.

3.5.1 Datasets

In this section, we present the different datasets we used for training and testing. One important aspect is the validation of our model's performance against the landmark dataset presented in section 3.3. The dataset consists of 139 landmarks derived from the INA thesaurus. Each landmark class contains approximately 200 images. However, due to the dataset's limited size, it is not sufficient for both training and testing purposes.

To address this limitation, we incorporate the Google landmark dataset, which is the most extensive dataset available to date, containing over 5 million images. To overcome computational constraints, we extract a subset from the Google landmark dataset, specifically selecting the landmark classes that align with the ones present in our constructed dataset (139 classes). The training set of the Google landmark dataset consists of approximately 400 images per class, providing a substantial amount of data for training our model.

For the testing phase, we utilize our constructed dataset (*cf.* Section 3.3) as a benchmark. In addition, to compare our work against previous state-of-the-art methods, we also include the Paris6k [28] dataset. This dataset comprises 6,000 images of landmarks in Paris.

While there is a domain shift between the training and test datasets, we are still able to conduct zero-shot evaluation using a recognition-by-retrieval process without employing finetuning on the new classes. This approach allows us to evaluate the performance of our model on the benchmark datasets without explicitly adapting it to the specific test classes.

3.5.2 Implementation details

To train the model, we apply transfer learning on ResNet-101, which was pre-trained on the ImageNet dataset. We remove the final fully connected layer and employ this pre-trained network to improve efficiency. The model undergoes training for 20 epochs using the Adam optimizer. We set the learning rate to $1e-6$ and the contrastive loss margin α to 0.85. Post-processing involves reducing the dimensionality of the global descriptor to $d = 256$, accomplished through PCA learned on the sample from the Google dataset. During training, all images are resized, to a size of 224×224 . During retrieval, we set the number of k nearest neighbors to 10.

3.5.3 Model evaluation on our dataset

In this section, we provide a comprehensive evaluation of our model's performance on our dataset. To assess the effectiveness of our model, we employ two widely-used evaluation metrics: Mean Average Precision (MAP) for the image retrieval task and accuracy for the recognition task.

MAP, or Mean Average Precision, is a commonly used metric in information retrieval. It measures the quality of ranked retrieval results by considering the precision and recall at various cut-off points. In the context of image retrieval, MAP evaluates the ability of our model to accurately rank and retrieve relevant images based on a given query. A higher MAP score indicates superior performance in retrieving relevant images.

For the image retrieval task on our constructed dataset, our model achieves a MAP score of 87.65%. This suggests that our model effectively ranks and retrieves relevant images for a given landmark query, showcasing its proficiency in recognizing landmarks within our dataset.

Additionally, we evaluate the recognition task on our dataset using the accuracy metric, which measures the proportion of correctly classified images. Our model achieves an accuracy score of 89.84%, further confirming its strong recognition capabilities.

3.5.4 Comparison with state-of-the-art

In this section, we compare our approach to previous state-of-the-art models on the widely-used benchmark dataset, Paris6k [28]. It is important to highlight that our model leverages a zero-shot evaluation approach, which holds significance in terms of evaluating model performance across different data distributions.

Zero-shot evaluation refers to assessing the performance of a model on a dataset without fine-tuning or adapting it specifically to that dataset. This approach ensures that the comparison between different models is unbiased, as it eliminates the possibility of models benefiting from specific dataset adaptations.

To ensure a fair comparison, we only consider methods that do not perform fine-tuning on the Paris6k dataset. This criterion ensures that the models being compared have not gained an unfair advantage by specifically adapting to the characteristics of the Paris6k dataset. The results are presented in Table 3.2.

Table 3.2. Comparison with state-of-the-art methods in landmark recognition and retrieval tasks on Paris6k dataset.

Method	MAP	Accuracy
VLAD-CNN [22]	-	58.3%
Crow [34]	79.7%	84.8%
BLCF [35]	82.0%	84.8%
R-MAC [23]	83.0%	86.5%
Ours	90.34%	91.48%

The selected models leverage pre-trained CNNs, typically trained on datasets like ImageNet, to enhance their performance in landmark recognition and retrieval tasks. Each model employs unique techniques to achieve this goal. Crow adopts a strategy of using cropped ROI images as input for the CNN. VLAD-CNN and R-MAC extract local features from intermediate layers of CNNs (GoogleNet and ResNet, respectively) and apply compact encoding/pooling techniques. R-MAC incorporates a modified version of the MAC feature aggregation method. BLCF utilizes VLAD [18] for feature aggregation. However, our model stands out by employing additional methodologies to further enhance performance. Through Siamese learning with contrastive loss and PCA whitening and dimensionality reduction techniques, we were able to surpass previous state-of-the-art methods by 7.34% in terms of MAP and 4.98% in terms of accuracy. These improvements highlight the efficacy of our approach and contribute to advancing the field of landmark recognition and retrieval.

3.5.5 Qualitative results

In this section, we present the qualitative results of our landmark recognition and retrieval work on the dataset we constructed (Figure 3.12) and Paris6k dataset (Figure 3.13). By showcasing the top 10 nearest images to the selected query, we demonstrate the effectiveness of our model in accurately retrieving relevant images.

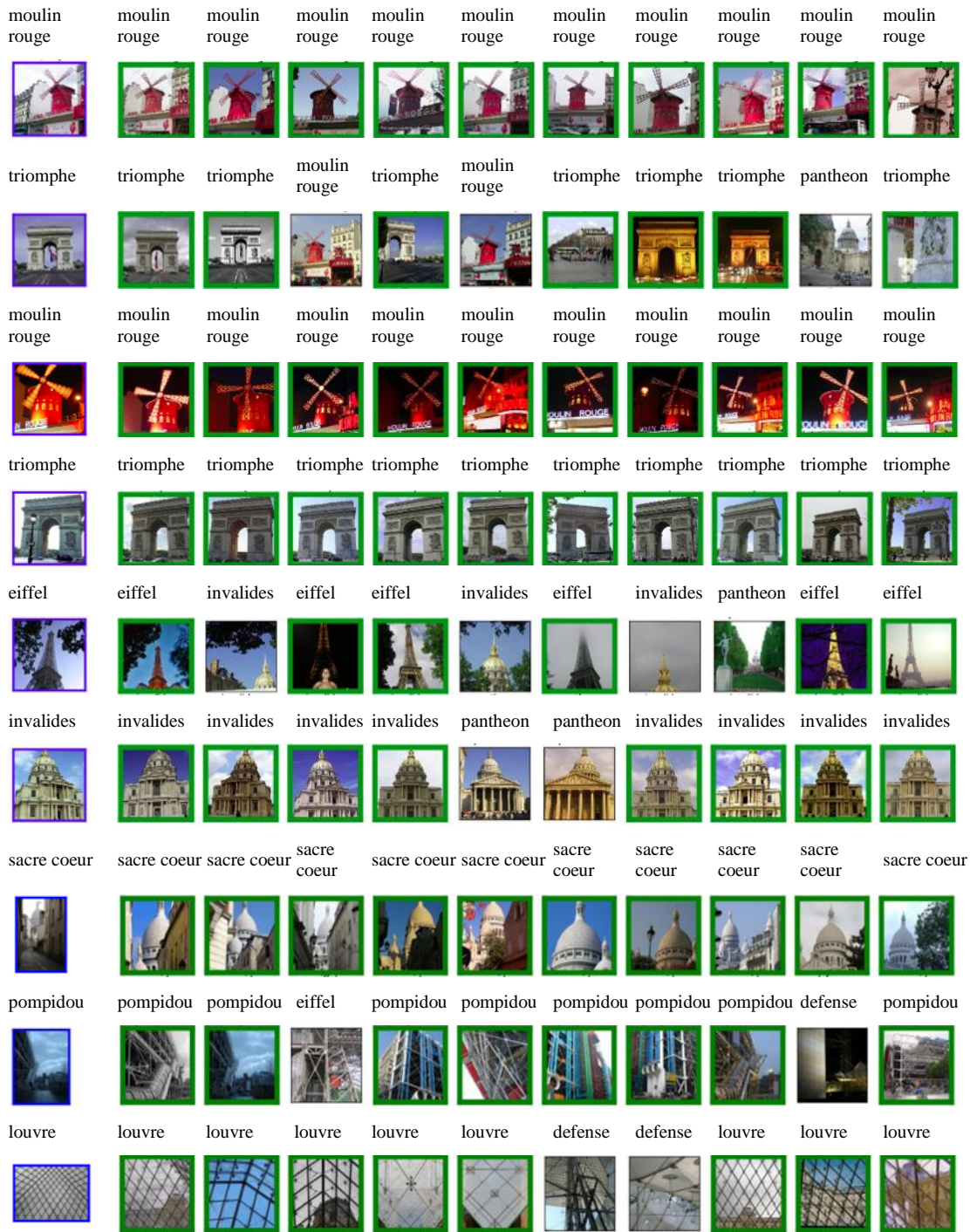


Figure 3.12. Retrieval examples from our dataset.

Our model exhibits robust performance in various scenarios, including accurate retrieval for both indoor and outdoor landmarks. Even in the presence of distractors, such as individuals in the images, our model successfully identifies the correct label of the landmark. This capability highlights the model's ability to focus on and recognize the main subject of interest amidst potential distractions. Furthermore, our model demonstrates its adaptability to different illumination situations. It successfully handles images captured in challenging conditions like cloudy or rainy days, as well as images taken at night. This capability showcases the model's robustness and effectiveness in dealing with varying lighting conditions commonly encountered in real-world scenarios.

In certain cases, our model encounters challenges when dealing with close-up shots, particularly when distinguishing between similar images. For instance, when presented with a close-up of the top of the Louvre, as shown in Figure 3.13 example 5, the model may mistakenly identify it as the foot of the Eiffel Tower. This difficulty can be attributed to the limited contextual information available in close-up shots. The model relies heavily on visual cues and patterns to identify landmarks, and when the image lacks broader context, it becomes more challenging to accurately differentiate between similar details. This issue is not exclusive to the model; even for humans, discerning such subtle differences can be challenging, especially when the images share common materials or intricate details. Additionally, the model's performance can be influenced by the availability and quality of training data, as close-up shots are not sufficiently covered.



Figure 3.13. Retrieval examples from Paris6k dataset.

3.6 Conclusion and future works

This chapter has presented a novel framework for landmark recognition, leveraging a recognition-by-retrieval technique that allows us to handle dataset changes and conduct zero-shot evaluation without requiring additional model finetuning. The ResNet architecture serves as the foundation of our retrieval system, enabling us to extract local features and aggregate them using MAC feature aggregation. Training our model entails utilizing contrastive loss through a siamese network. Additionally, we incorporate dimensionality reduction and whitening techniques to learn more robust features.

To evaluate the effectiveness of our model, we generated a landmark dataset tailored to the specific requirements of documentalists, using an automated technique that draws from the INA thesaurus for dataset class derivation. Due to computational limitations, we trained our model on a subset of the Google landmark dataset.

The results obtained demonstrate the strong performance of our model on our dataset, achieving an MAP of 87.65% for the retrieval task and an accuracy of 89.8% for the recognition task. Furthermore, we conducted a comprehensive comparison with state-of-the-art methods using the Paris 6k dataset, surpassing them with a 7.34% higher MAP and 4.98% higher accuracy.

For future work, we have identified several axes to explore. Firstly, we intend to further improve the performance of our model by expanding the training data and incorporating more diverse landmark images. Additionally, we plan to investigate the application of our framework to real-time landmark recognition scenarios. Furthermore, exploring transfer learning techniques and investigating the impact of different network architectures could provide valuable insights for enhancing the overall performance and scalability of our system.

4 CONTRIBUTION TO THE SCENE SEGMENTATION PROJECT

Abstract: This chapter presents our contribution to the scene segmentation project by focusing on two essential tasks: indoor-outdoor place recognition and field of view shot type identification. These tasks provide highly useful cues that can be exploited for effectively clustering different shots into coherent scenes. Different other applications, such as video analysis, archive indexing, and TV program metadata indexing, can also benefit from such features. Throughout this chapter, we describe the methodologies employed for both tasks, including the selection of appropriate features, training of machine learning models, and the evaluation of their performances. The experimental results obtained on diverse datasets demonstrate the effectiveness and robustness of the proposed methods.

Keywords: Scene classification, Field of view shot type, place recognition.

4.1 Introduction

In this chapter, we present our contribution to the scene identification project, which is an integral part of a broader project dedicated to video segmentation purposes. In the TV broadcast community, video segmentation holds a significant importance for various reasons. One key aspect is archive indexing, where videos are indexed shot by shot. Traditionally, editors manually input time codes for each shot. However, manual shot boundary indexing is often unreliable and prone to errors, which can hinder efficient retrieval of content. Another application where video segmentation plays a crucial role is ad insertion. To enhance the viewer experience, advertisements need to be inserted at temporal locations that align with content discontinuity, such as the end or start points of action plots. This requires accurate segmentation of videos into shots and scenes, enabling precise identification of suitable insertion points. Figure 4.1 illustrates the ad insertion framework developed within the AI-TV joint laboratory, to which we have delivered our contributions for integration. For each shot, the framework identifies the type of environment: indoor/outdoor, the shot type (long-shot, close-up...) and the place category along with its estimated probability (restaurant, bar, office...). The related metadata are used for story identification as explained in the following section.



Figure 4.1. The AI-TV ads insertion framework

4.2 Scene segmentation overview

Video segmentation involves dividing a video into distinct units called shots. Each shot represents a continuous sequence of frames captured by the camera without interruption. Once the shots are identified, the system cluster them into story units called scenes. The scene identification is based on various semantic and visual criteria. The complete methodology is described in details in [36]. One of the components considered is notably related to the contextual similarity between adjacent, successive shots (s_n, s_{n+1}). The contextual similarity, denoted by $Sim_{context}(s_n, s_{n+1})$ takes into account the number of common places between the two successive video shots considered and their corresponding recognition probabilities, and is defined as:

$$Sim_{context}(s_n, s_{n+1}) = \frac{1}{L} \sum_{i=1}^k \alpha(s_n) \cdot p_i^n + \beta(s_{n+1}) \cdot p_i^{n+1} \quad (4.1)$$

where p_i^n denote the i^{th} place recognition probability in shot s_n , k is the total number of common locations recognized in shots s_n and s_{n+1} , L is the number of shots, while parameters α and β control the influence of the video camera filming type. For wide, long and medium shots the values for α and β are fixed to 1, while for all the others shots the value for α and β is 0.5.

Our contributions to this project are notably related to the two different items involved in equation (4.1). A first one concerns the identification of both indoor and outdoor places that are present in the video scenes (section 4.3). The second one is related to the classification of the field of view of each shot (section 4.4). In both cases, existing state of the art techniques have been adopted.

4.3 Place recognition

Place recognition refers to the task of identifying and recognizing specific locations or places (e.g. restaurant, museum, studio) in an environment based on visual cues. Place recognition plays a vital role not only in scene identification but also serves as crucial metadata in the process of archive indexing, as detailed in Section 2.3.6. To this purpose, we have adopted the GoogLeNet architecture [37].

4.3.1 Network architecture

The GoogLeNet model introduce the concept of the *inception module*, which uses relatively dense components to approximate the optimal local sparse structure. Within this framework, the network consists of six convolution layers and one pooling layer, followed by six inception modules. Finally, the characteristic parameters are transferred through a fully connected layer. A ReLU nonlinear activation function is added in each layer to reduce the probability of gradient disappearance and improve the speed of the backpropagation calculation. The maximum pooling layer can reduce the error of the estimated mean deviation caused by the parameter error of the convolution layer and retain more texture features.

The Inception module (Figure 4.2) uses an asymmetric convolution kernel to replace the conventional convolution kernel and thus reduce the computational burden.

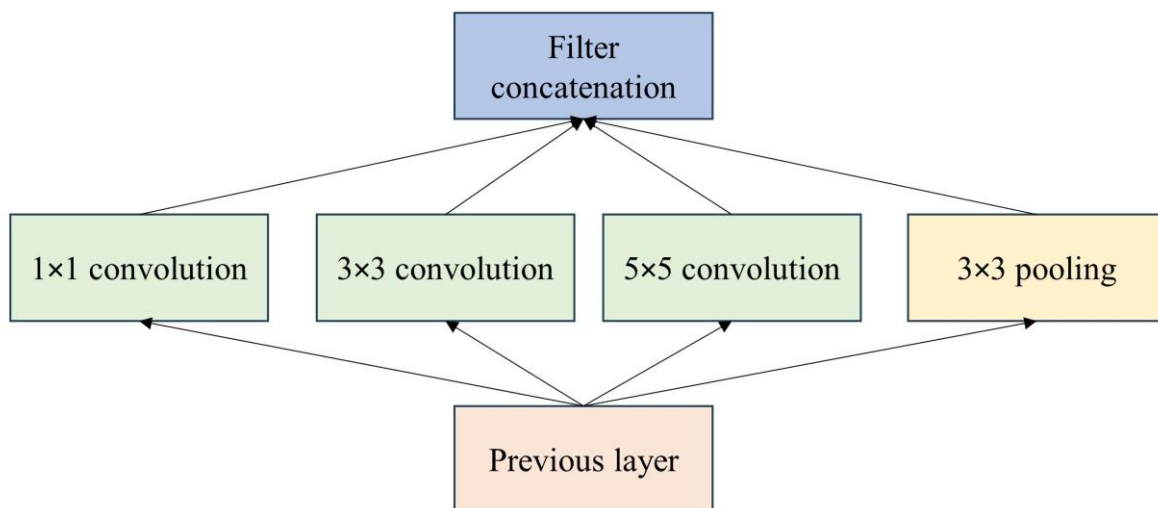


Figure 4.2. Inception block.

4.3.2 Experimental setup

We initialize the model with pre-trained weights on ImageNet. For finetuning, we have used the Adam optimizer with a learning rate of $1e-3$. We train the model for 20 epochs on Places365 [38] dataset which is specifically designed for scene recognition tasks. It consists of a vast collection of 10 million images, covering 434 different scene classes. The dataset offers two versions: Places365-Standard and Places365-Challenge2016. The Places365-Standard version comprises 1.8 million training images and 36,000 validation images, representing 365 scene classes. On the other hand, the Places365-Challenge2016 version expands the training set to 6.2 million additional images, including 69 new scene classes, resulting in a total of 8 million training images from the 434 scene classes. For the experiments conducted in this project, the Places365-Standard dataset has been utilized.

4.3.3 Model evaluation

For evaluation, we have considered the Top-1 and Top-5 (the percentage of images where the ground truth is among the top 5 predicted labels) accuracy rates. The reason behind this is that the scenes are inherently multi-labels in terms of their semantic description and the Top-1 accuracy can be an ill-defined measure in this case. The learned model achieves 55,6% Top-1 accuracy and 85,66% Top-5 accuracy on the validation dataset of Places365. Finally, we assign a label indoor/outdoor to each shot based on the majority of predicted places. We also evaluate our approach on random samples from programs of France TV to assess the capability of generalization of the model on unseen data. Figure 4.3 presents some examples. The model showcases its ability to distinguish diverse scenes and categorize environments as indoor or outdoor accurately. However, the confidence levels of its predictions can vary widely based on the complexity of the image. While the model performs commendably, the lower confidence in some of the predictions implies room for improvement in model certainty, which could potentially enhance both Top-1 and Top-5 accuracies. Overall, the model's performance is promising, yet further refinement on specific France TV data may yield even better results.



top-1 : television studio (0.350)
top-2 : booth (0.049)
top-3 : legislative chamber (0.048)
top-4 : conference center (0.034)
top-5 : arena/performance (0.03)

Type of environment: indoor



top-1 : booth (0.213)
top-2 : reception (0.211)
top-3 : legislative chamber (0.176)
top-4 : beauty salon (0.131)
top-5 : conference center (0.059)

Type of environment: indoor



top-1 : street (0.262)
 top-2 : crosswalk (0.089)
 top-3 : highway (0.087)
 top-4 : embassy (0.076)
 top-5 : parking lot (0.06)

Type of environment: outdoor

Figure 4.3. Examples of the top-5 predictions from France TV content. The number beside indicates the prediction confidence.

4.4 Field of view shot detection

Within the context of video analysis, the shot type refers to the degree of visual proximity between the camera and the scene represented in a given frame. Different shot types provide different perspectives and convey different levels of detail or focus on the subject. In this section, we present an approach for categorizing video shots into six distinct shot types based on the proximity of the camera to the objects. These shot types are the following (Figure 4.4):

- *Extreme Wide Shot (EWS)*: In this case, the camera captures a wide view of the scene, showing a significant portion of the environment or location. The focus is on providing context and establishing the setting rather than specific details.
- *Large Shot (LS)*: The camera is positioned at a considerable distance from the subject, capturing a broader view compared to other shot types. While it includes more of the scene than close-up shots, it still maintains some distance, allowing for a wider perspective.
- *Medium Shot (MS)*: In a medium shot, the camera is closer to the subject, resulting in a framing that captures the subject from the waist up or from the knees up. It offers a moderate level of detail while maintaining a wider view of the surroundings.
- *Medium Close-Up (MCU)*: The camera is positioned closer to the subject in a medium close-up shot. This shot focuses on capturing the subject from the chest or shoulders up, providing a more detailed view while still maintaining some context of the surrounding environment.
- *Close-Up (CU)*: In a close-up shot, the camera is placed very close to the subject, emphasizing facial expressions, specific details, or objects of interest. It typically frames the subject from the neck or shoulders up, creating an intimate and focused view.
- *Extreme Close-Up (ECU)*: An extreme close-up shot involves the camera being extremely close to the subject, capturing only a small portion or detail of the subject. This shot type is often used to highlight specific features, emotions, or objects in great detail.

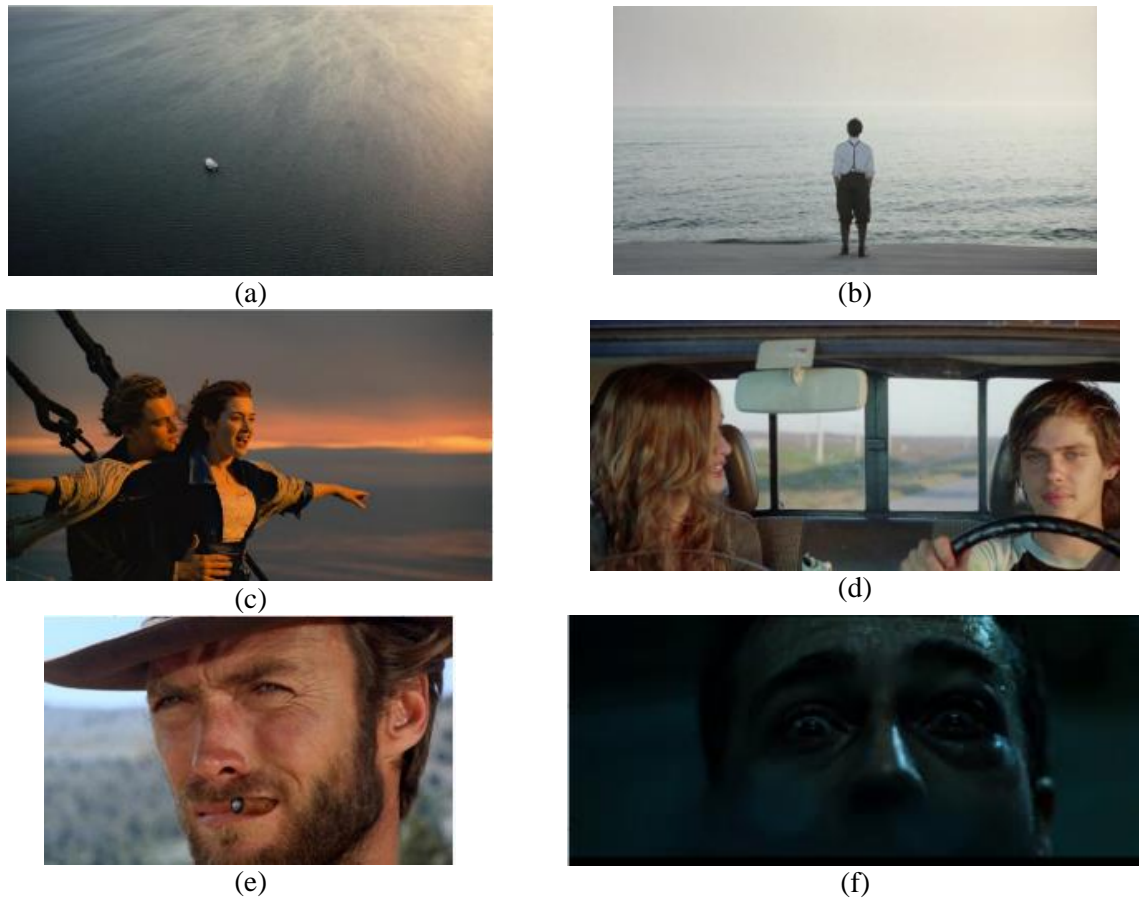


Figure 4.4. Basic field of view shot types. a) EWS; b) LS; c) MS; d) MCU; e) CU; f) ECU.

4.4.1 Experimental setup

To train our model for shot type identification, we have utilized the publicly available Film-grab database [39], which consists of a total of 5,505 images for training and 600 images for validation. We employed the Resnet-50 [27] network architecture pre-trained on Imagenet. For finetuning, we have employed the cyclical transfer learning technique [40]. Cyclical learning involves training the model in multiple cycles, with each cycle involving an adjustment to the training data or model parameters. In this experiment, we initially train the model with lower-resolution images and then progressively train on higher resolution. Training on lower-resolution images first can help the model learn more general and coarse-level features that are relevant across different resolutions. As the model becomes proficient at recognizing and extracting these features, it provides a solid foundation for subsequent training on higher-resolution images, where more fine-grained details and local features become important. This progressive learning approach can lead to improved performance in capturing both global and local characteristics of the data. At each stage, the network is first trained with a learning rate of $1e-3$ using ADAM as optimizer and finetuned with a learning rate of $1e-6$.

4.4.2 Model evaluation

The model demonstrates a high accuracy rate of 91% on the validation set. To gain further insights and enhance the interpretability of the results, we analyze the confusion matrix, presented in Figure 4.5. From the matrix, we observe that the model performs exceptionally well in classifying extreme wide shots, close-ups, and extreme close-ups. However, some instances of confusion arise between long shots and medium shots. It is worth noting that the labeling process is subjective and performed manually, which could account for the disparity in predictions and occasional misclassifications.

EWS	0.94	0.03	0.01	0	0.01	0.01
LS	0.01	0.82	0.14	0	0.02	0.01
MS	0	0.03	0.89	0.07	0	0.01
MCU	0	0	0.06	0.92	0.02	0
CU	0	0	0	0.02	0.94	0.04
ECU	0	0.02	0.01	0	0.02	0.95
	EWS	LS	MS	MCU	CU	ECU

Figure 4.5. Confusion matrix of the test set

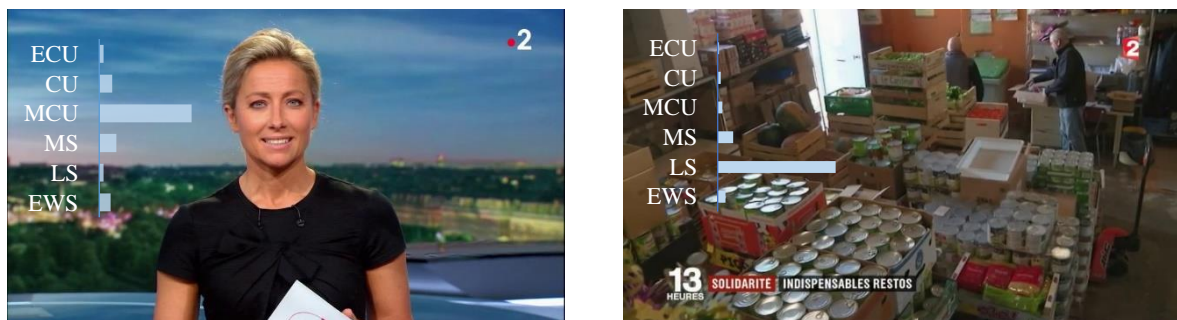


Figure 4.6. Examples of field of view shot recognition on France TV content. Horizontal bars indicate the prediction confidence.

Since the dataset was extracted from Hollywood movies we evaluate also the model's performance against France TV content using a sample of 100 images (Figure 4.6). Some additional prediction examples are presented in Figure 4.7. The obtained error rate of 11% shows that the network detects fundamental features and can generalize well the results.



GT:MS, P:MS



GT: CU, P: CU



GT: EWS, P: LS



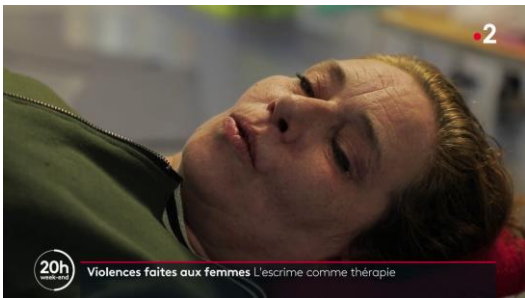
GT: LS, P:LS



GT: MS, P:MS



GT: MCU, P: CU



GT : CU, P : CU



GT : EWS, P :EWS

Figure 4.7. Examples of Field of view shot type prediction. (GT: Ground Truth, P: Prediction)

4.5 Conclusion

In this chapter, we have presented our contribution for the scene identification project. We used state-of-the-art techniques for place recognition and field of view shot detection. The models achieved high accuracy rates on the validation datasets from Places365 and FilmGrab respectively. Evaluation on France TV content demonstrated the model's ability to generalize well, with a low error rate of 11%. Overall, our contributions improve scene identification and enhance the efficiency of video segmentation tasks.

For future work, we plan to employ the datasets that have been manually annotated by documentalists to construct a specialized dataset specifically for France TV. The rich archives of France TV are incredibly valuable due to their diverse range of genres, cinematographic styles, and cultural contexts. This diversity is a source of unique challenges that are tailored to the needs of the industry. Leveraging this comprehensive testing ground can facilitate key insights that further refine and enhance our methodologies.

5 CAMERA MOTION CATEGORIZATION

Abstract: The automatic estimation of the various types of camera motion (e.g., traveling, panning, rolling, zoom...) that are present in videos represents an important challenge for automatic video indexing. Previous research works are mainly based on motion vectors/optical flow estimation and analysis. In this chapter, we propose a different, deep learning-based approach that makes it possible to classify the videos according to the type of camera motion. The proposed method is inspired from action recognition approaches and exploits 3D convolutional neural networks with residual blocks. The performances are objectively evaluated on challenging videos, involving blurry frames, fast/slow motion, poorly textured scenes. The accuracy rates obtained (with an average score of 94%) demonstrate the robustness of the proposed model.

Keywords: Camera motion classification, deep learning, Resnet, 3D CNN.

5.1 Introduction

The camera motion pattern plays a crucial role in the field of TV broadcasting, serving various purposes within the video content production process. This process involves a diverse range of professionals, each contributing with their specific expertise. During video shooting, operators capture a vast array of shots, employing different types of camera motions to enhance the visual experience. Following the shooting phase, the selection of the most suitable shots falls upon the video editors, who employ aesthetic, artistic, and operational criteria. Here, the camera motion type emerges as a critical aspect that demands careful consideration.

Unfortunately, the information regarding camera motion is often lost during the editing process, requiring video editors to manually search and identify the most appropriate shots. Consequently, the absence of an automated motion type classification tool significantly hinders the efficiency of the video editing workflow.

Beyond video editing, camera motion classification also holds importance in archive indexing, browsing, and retrieval processes. Within these contexts, documentalists extensively engage in tagging key components of a video to ensure accurate classification within the archive and enable effective retrieval and re-use strategies. The camera motion type is thus a mandatory field to be specified by documentalists.

In this chapter, we explore the use of a 3D CNN model for camera motion classification. Like any deep learning-based approach, the related constraint relies on the availability of an important volume of labeled samples, which are necessary for setting up a successful training process. However, currently there is no such publicly available data set in the literature, because of the tedious and expensive manual annotation process required to constitute such a corpus.

The fundamental question that we have to solve is then the following: how can we learn a deep CNN model dedicated to camera motion categorization purposes, with a limited amount of training data?

In order to solve this problem, the first contribution consists of a transfer learning-based camera motion classification method. The originality of the approach comes from the fact that the initial learning is performed on a data set that is completely different, in terms of targeted classes, from our camera motion categories. More precisely, we use for training the Kinetics [38] action corpus, which consists of 400 action categories. Such a corpus has nothing to do with our purpose. However, we claim that the derived feature maps capture essential, salient spatio-temporal cues that can be exploited for camera motion classification. A fine-tuning is then applied on a dedicated camera motion data set with a reduced number of items. Our second contribution notably concerns a semi-automatic method that makes it possible to construct a reliable camera motion dataset from general public videos with a minimum amount of human intervention.

Finally, the third contribution concerns the creation of a camera motion evaluation dataset. The corpus includes highly challenging videos, acquired in real-life conditions with professional cameras and at various resolutions. It allowed us to assess the robustness and power of generalization of the proposed technique, which yields an average accuracy rate of about 94%. To the very best of our knowledge, our work presents the first data-driven solution to characterize camera motion in videos using a deep 3D convolutional neural network.

The remainder of this chapter is organized as follows. In Section 5.2, we present the various forms of camera motion employed in cinematography, aiming to establish a foundational understanding. We then proceed with a state of the art review. In Section 5.3.1, we notably consider and analyze existing camera motion estimation algorithms. As our research also derives inspiration from advances in the field of video action recognition, in Section 5.3.2 we analyze existing work on this subject. Section 5.4 details the proposed methodology, with retained architecture as well as training and validation datasets and

protocols. The experimental results obtained are presented and discussed in Section 5.5. Finally, we summarize our findings and draw our future perspectives in Section 5.6.

5.2 Types of camera motion

Camera movement refers to the intentional change in the position, orientation, or focal length of a camera during the process of capturing video footage. It adds a dynamic element to the visual presentation and significantly impacts the viewer's perception and engagement with the content.

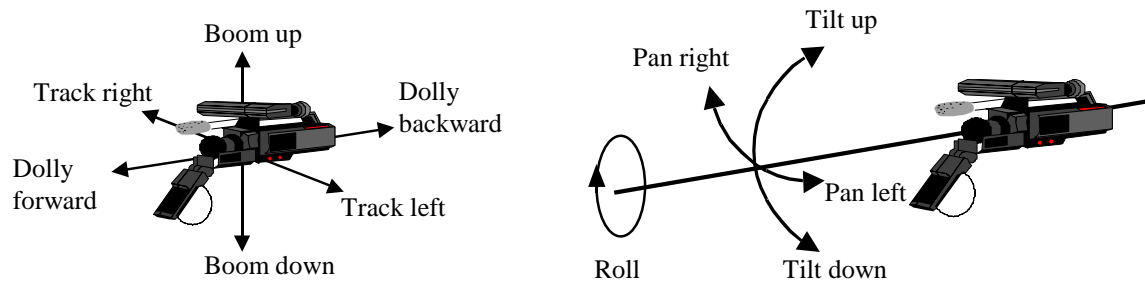


Figure 5.1. Different types of camera movement

The different types of camera movements (Figure 5.1) most often used in video production are the following:

- *Pan*: A pan involves horizontally rotating the camera on its axis from a fixed position. It produces a sweeping motion across the scene, allowing the camera to cover a wider area. Panning is often used to follow a subject's movement or to reveal elements from one side of the frame to another.
- *Tilt*: Tilt refers to the vertical rotation of the camera on its axis while maintaining a fixed position. It allows the camera to look up or down, capturing scenes from different angles. Tilt movements are often used to emphasize height, depth, or to follow the movement of subjects located at different vertical positions.
- *Tracking*: Tracking involves smoothly moving the camera horizontally. This dynamic movement adds a sense of fluidity and perspective to a scene, allowing the audience to follow a subject's lateral movement or explore the environment with a seamless visual experience.
- *Dolly*: Dolly involves physically moving the camera towards or away from the subject. It creates a smooth and fluid motion, enabling the camera to follow the subject's movement or change the perspective dynamically. Tracking shots are commonly used to add a sense of depth and immersion to the footage.
- *Zoom*: Zooming is the adjustment of the camera's focal length, either by using an optical zoom lens or changing the zoom digitally. Zooming in increases the magnification, bringing the subject closer, while zooming out decreases the magnification, capturing a wider field of view. Zooming can be used to create a sense of intimacy, highlight details, or establish the context of a scene.
- *Crane or Boom*: Crane or boom shots involve the camera being mounted on a crane or boom arm to achieve sweeping vertical or horizontal movements. These shots offer a broad range of motion possibilities, such as high-angle or low-angle shots, and are often used to create dramatic or visually striking effects.

The different camera movements offer a wide range of creative possibilities for capturing video footage. Skilled cinematographers and camera operators strategically employ these movements to enhance storytelling, create visual interest, and evoke specific emotions or moods in the audience.

5.3 Related work

Camera motion estimation is commonly employed as a key building block, contributing to the advancement of several fields including autonomous navigation, robotics, and augmented reality. Traditional approaches are based on the analysis of the motion vectors between successive frames, which serve to determine higher-level parametric motion models, such as affine [41], simplified affine [42], [43] planar perspective or homographies [43] that can globally describe the camera motion. Concerning the motion vectors, which represent the essential feature to be considered, they can be determined by simple block matching techniques, dense optical flow approaches or tracking of interest points [44]. The camera motion type is then estimated by comparing the parameters, or corresponding motion vectors, to predefined thresholds.

5.3.1 Estimation of motion vectors

Existing motion vector estimation techniques can be categorized into two methodological families, including feature-based and appearance-based approaches. The former relies on the extraction and tracking of salient visual features, while the latter focuses on analyzing the appearance information directly from the image sequences.

5.3.1.1 Feature-based approaches

The feature-based approaches have been widely employed in camera motion estimation [45]–[50]. Most of the time, they rely on the extraction of a set of interest points, with the help of well-known techniques such as the Harris or Harris-affine corner detectors [51]. Such methods ensure a good repeatability of the detection in successive frames. Furthermore, the detected interest points are described with the help of local invariant features such as SIFT (*Scale-Invariant Feature Transform*) [10], SURF (*Speeded-Up Robust Features*) [15] and FAST (*Features from Accelerated Segment Test*) [52] descriptors. Once the features are extracted for each pair of consecutive frames, the next step is to match or track these features in order to estimate the motion vectors. A direct way to determine matches is to compute the Euclidean distance between the corresponding descriptors and select as matches the features with smaller distances. The RANSAC (*Random Sample Consensus*) algorithm [53] can be also used in combination with such approaches in order to robustly estimate a global parametric motion model, minimizing the influence of outliers.

5.3.1.2 Appearance-based approaches

The appearance-based approaches, also known as optical flow techniques [54]–[57] focus on measuring changes in the appearance and image intensity values. The principle consists in estimating a motion field based on an assumption of conservation of the luminance signal between consecutive frames.

There are two types of optical flow algorithms: dense and sparse. Dense optical flow algorithms, such as the Horn-Schunck algorithm [58] estimate the displacement at each pixel by incorporating global regularity constraints. On the other hand, sparse optical flow algorithms, such as the Lucas-Kanade method [59] calculate the displacement for a selected number of pixels in the image.

While dense optical flow algorithms eliminate the need of interest point extraction, they are more sensitive to noise when compared to sparse optical flow algorithms. Consequently, sparse optical flow algorithms are more adapted for applications where a robust displacement estimation is required. However, it is important to carefully choose the features, considering that pixels in regions with higher variance among neighbors will yield more reliable displacement estimation.

5.3.1.3 Discussion

The feature-based approach is commonly employed in textured environments such as rough and urban areas [57], [60]. However, this approach encounters challenges when dealing with texture-less or low-textured environments characterized by a single pattern, such as blank walls, clear blue skies, and featureless water surfaces [61]. In such scenarios, the feature-based approach proves to be inefficient due to the limited number of salient features that can be detected and tracked [54]–[56]. In contrast, the appearance-based approach exhibits greater robustness and superiority over feature tracking methods in low-textured environments [54], [62]. By using larger templates during the matching process, the appearance-based approach achieves a higher probability of successful matching between consecutive image frames. In certain scenarios, a hybrid approach [63] that combines both feature-based and appearance-based techniques proves to be the most effective solution. Such hybrid approaches leverage the tracking of salient features across frames while utilizing pixel intensity information from the entire or batches of images.

However, in real life videos the estimation of motion vectors faces various difficulties:

- *Lack of distinctive features:* Without well-defined and recognizable points, the tracking process becomes challenging or even infeasible.
- *Complex camera trajectories:* Advanced shots often involve complex camera trajectories, with varying speeds, angles, and directions of movement. Interest point tracking algorithms may struggle to accurately track points along such intricate trajectories due to the limitations of the tracking algorithm or the instability introduced by the camera movement itself.
- *Perspective changes:* The perspective changes can cause significant distortions and variations in the appearance of interest points, making it difficult for tracking algorithms to maintain a consistent matching tracking across frames.
- *Inherent camera shake:* The resulting vibrations or instability can introduce additional challenges for interest point tracking algorithms, leading to inaccurate or unstable tracking results.

Due to such difficulties, in our work, we have considered the motion vector-based traditional techniques uniquely to the purpose of constituting in a semi-automatic manner a video corpus that can further serve as a training dataset for advanced deep learning camera categorization approaches. Notably, our work draws inspiration from video action recognition methodologies.

Video action recognition has experienced a remarkable growth in the field of video understanding. Deep learning-based research has predominantly concentrated on this domain, with numerous techniques and methodologies developed to tackle the challenges of recognizing and interpreting human actions in videos. Most of the time, the video action recognition techniques rely on features extracted from 2D or 3D CNNs. Such approaches are reviewed in the following section.

5.3.2 CNNs for video action recognition

The field of action recognition has witnessed significant progress in recent years, driven by advancements in deep learning techniques. In particular, Convolutional Neural Networks (CNNs) have emerged as a powerful tool for extracting discriminative features from video data, enabling accurate

recognition and understanding of human actions. The state-of-the-art reveals two main families of CNN-based approaches.

The first one originated from the influential paper on Two-Stream Networks introduced in [64]. It involves incorporating an additional pathway to train a convolutional neural network on the optical flow stream, thereby capturing temporal information within a video. The remarkable success of this approach has inspired a considerable number of subsequent papers, including TDD [65], LRCN [66], Fusion [67] or TSN [68].

The second family of approaches revolves around the utilization of 3D convolutional kernels to effectively model temporal information in videos. Examples include I3D [69], R3D [70], S3D [71], Non-local [72], and SlowFast [73].

5.3.2.1 Two-stream networks

In the era of deep learning [74], there has been a growing interest among researchers to apply CNNs to video-related tasks. The DeepVideo technique [75] propose to employ a single 2D CNN model independently on each video frame. Authors explore various temporal connectivity patterns, including late fusion, early fusion, and slow fusion, to capture spatio-temporal features for video action recognition. However, it became evident that solely 2D CNNs lack the ability to effectively incorporate temporal information, rendering them insufficient for comprehensive video understanding. Thus, Simonyan *et al.* [64] introduced the concept of two-stream networks, illustrated in Figure 5.2, which includes two different analysis streams, temporal and spatial. The approach is inspired by the two-streams hypothesis [76], which suggests that the human visual cortex includes two distinct pathways: the ventral stream for object recognition and the dorsal stream for motion perception. In line with this assumption, the spatial stream is responsible for processing raw video frames to extract visual appearance details. On the other hand, the temporal stream utilizes a stack of optical flow images as input to capture motion information between consecutive video frames.

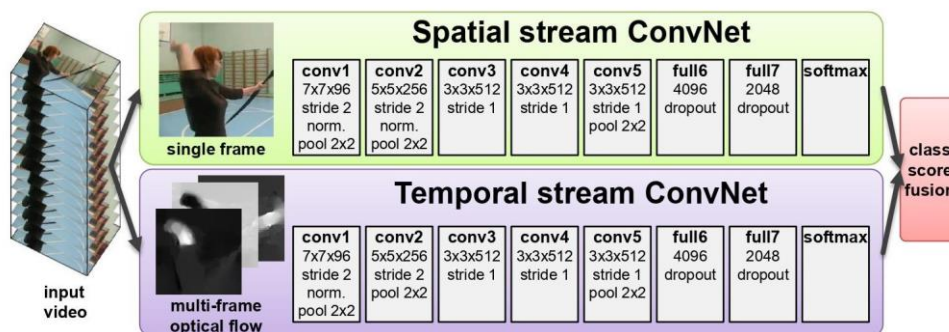


Figure 5.2. Workflow of two-stream network [64].

The fusion of spatial and temporal networks in a two-stream architecture presents a significant challenge known as spatio-temporal fusion. The late fusion, which involves a weighted average of predictions from both streams, is the simplest and most commonly used approach [64], [77]. However, researchers argue that early fusion, enabling earlier interactions between the networks, could benefit both streams during model learning. Among one of the first papers exploring early fusion techniques, let us mention the approach introduced in [67]. Here, authors investigate various spatial fusion methods (e.g., sum,

max, bilinear, convolution, concatenation), network layers for fusion, and temporal fusion approaches (e.g., 2D or 3D convolutional fusion in later stages of the network). Results showed that early fusion enhances both streams' ability to learn richer features, leading to improved performances when compared to late fusion.

Expanding on this research, Feichtenhofer *et al.* [78] extended the ResNet model [27] to the spatiotemporal domain by introducing residual connections between the two streams. They also proposed a multiplicative gating function [79] to improve the learning of spatio-temporal features in residual networks. Concurrently, the approach introduced in [80] exploits a spatio-temporal pyramid to perform hierarchical early fusion between the two streams.

Let us underline that optical flow is crucial in two-stream networks, but the pre-computation of optical flow is computationally demanding and resource-intensive, making it challenging for large-scale training and real-time deployment.

As an alternative, videos can be conceptualized as 3D tensors with two spatial dimensions and one temporal dimension. This observation has led to the conception of 3D CNNs, described in the following section.

5.3.2.2 3D Convolutional Neural Networks

In their seminal work [81], Ji *et al.* introduce the concept of 3D CNNs for action recognition. However, the initial approach was not sufficiently deep to fully exhibit its potential for action recognition purposes. Tran *et al.* [82] extended this work by proposing a deeper 3D network called C3D, following the modular design of [37], similar to a 3D version of the VGG16 network. Although the C3D's performance on standard benchmarks was not fully satisfactory, it demonstrated strong generalization capabilities and served as a versatile feature extractor for various video analysis tasks [83].

Let us observe that training efficient 3D networks is challenging due to the need for large-scale datasets and the time-consuming nature of training, resulting in the dominance of 2D CNN-based two-stream networks for video action recognition during 2014-2017.

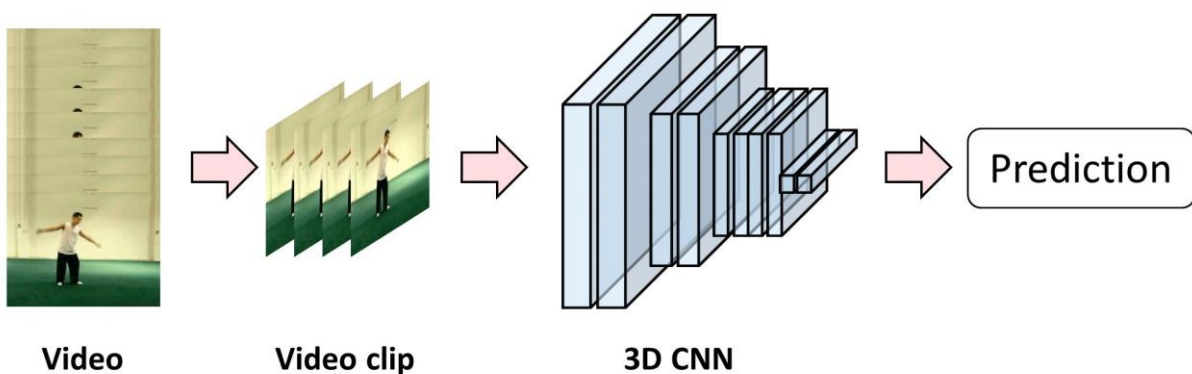


Figure 5.3. Workflow of 3D CNN.

The situation changed in 2017 with the introduction of the I3D model [69]. I3D employs stacked 3D convolutional layers to process video clips, typically consisting of 16 or 32 successive frames. Notably, I3D adapts mature image classification architectures for 3D CNNs and uses a method from [77] to

initialize model weights by inflating pre-trained 2D model weights. This approach allowed I3D to overcome the need of training 3D CNNs from scratch. This advancement propelled video action recognition to new heights, with 3D CNNs emerging as top performers across various benchmarks.

In an analogous manner, the ResNet3D (also known as R3D) model [70] adopts the concept of 2D ResNet [27] by replacing all 2D convolutional filters with 3D kernels. The approach aims to leverage the success of 2D CNNs on ImageNet by utilizing deep 3D CNNs alongside large-scale datasets. Building upon the idea of ResNeXt [84], Chen *et al.* [85] introduced a multi-fiber architecture that divides a complex neural network into a set of lightweight networks (fibers). This design facilitates information exchange among the fibers, thereby reducing computational costs. Drawing inspiration from SENet [86], STCNet [87] proposed the integration of channel-wise information within a 3D block. This integration allows to capture the correlation between spatial and temporal channels across the entire network.

To simplify the training of 3D networks, P3D [88] and R2+1D [89] employ the concept of 3D factorization. Specifically, a 3D kernel, such as $3 \times 3 \times 3$, can be decomposed into two separate operations: a 2D spatial convolution, like $1 \times 3 \times 3$, and a 1D temporal convolution, like $3 \times 1 \times 1$. The P3D and R2+1D approaches differ in the way they arrange the factorized operations and formulate each residual block. Another approach that follows this concept is the so-called trajectory convolution [90], which incorporates deformable convolution for the temporal component to better handle motion dynamics.

5.3.2.3 Discussion

Given these considerations, we have chosen to use 3D convolutional neural networks (CNNs) for camera motion estimation. While "factorized" networks like P3D or R2+1D offer elegant solutions to reduce complexity during both training and inference phases, they may not capture temporal dependencies as effectively as the R3D model. Furthermore, their performance tends to be lower when confronted with datasets containing challenging temporal variations [70], which is a crucial aspect of camera motion estimation. It is also worth noting that our solution is intended to be implemented offline rather than for real-time applications. Therefore, the slight ameliorations in efficiency achieved by "factorized" networks do not present a major problem in our specific case. Therefore, based on our observations and evaluations, we have decided to adopt the R3D model.

5.4 Proposed methodology

To the best of our knowledge, there is no publically available dataset accessible that can be used for training. Therefore, we have established a semi-automatic process to generate a custom dataset from YouTube videos. To assess our approach objectively, we have also created a validation dataset, gathered under various acquisition conditions.

5.4.1 Network architecture

For the camera motion estimation, we employ a 3D version of the Residual Network (ResNet) architecture, so-called 3D-ResNet [70]. The model incorporates shortcut connections between layers, which help in mitigating the issue of vanishing gradients during the training phase, particularly beneficial for deeper networks. Mathematically, it signifies that instead of learning an original mapping $H(x)$, a Resnet layer approximates the residual mapping $F(x) = H(x) - x$.

We have extended the ResNet-34 model by incorporating 3D convolutions into its architecture. This adjusted structure involves an initial convolutional layer of dimensionality $7 \times 7 \times 7$, subsequently followed by a max-pooling layer (see Figure 5.4). This sequence enables the extraction of robust features

while significantly reducing the spatial dimensions of the input data, thereby improving computational efficiency.

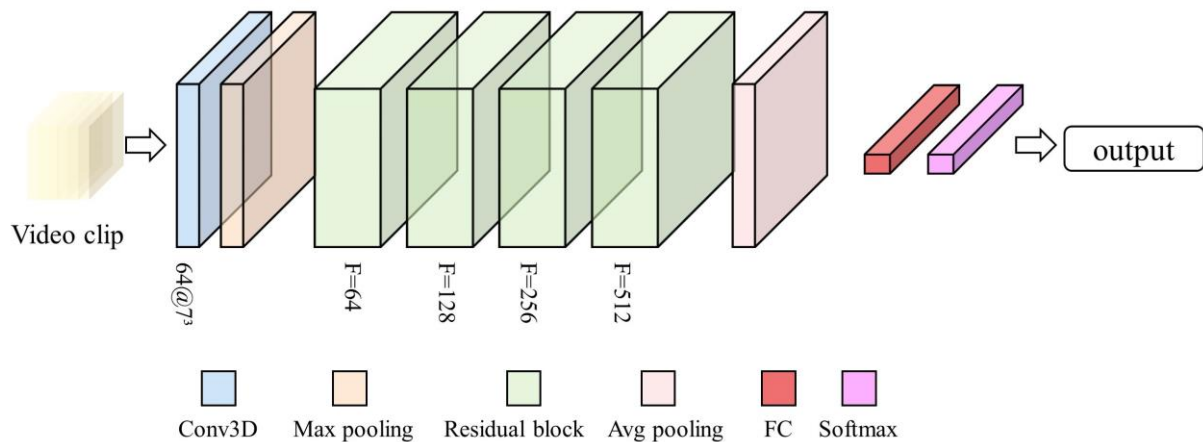


Figure 5.4. Illustration of the adopted 3D CNN. The notation $F@H^3$ means F filters of size $H \times H \times H$.

The next part of our architecture includes four residual units. We leverage the basic ResNet block characterized by type 'A' shortcut connections (Figure 5.5) [27]. This choice involves utilizing zero-padding to ensure that the shortcut paths and the main paths have the same dimensions, permitting seamless addition. Inside these residual units, the blocks are constructed of a sequence of 3, 4, 6, and finally 3 convolutions respectively. The diverse composition of convolutional layers within each unit allows the network to learn a wide array of features at varying levels of complexity. We also spatially down-sample the input data within the third to fifth bottleneck. Here, we employ a stride of $s=2$, effectively reducing the size of the input.

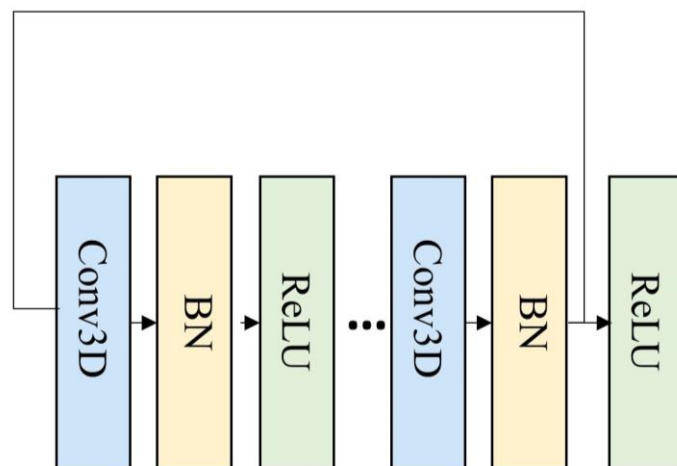


Figure 5.5. Skip connection employed in the network.

Following the residual units, we employ a global average pooling layer. This layer simplifies the output by calculating the mean of each feature map, preserving the depth dimension but reducing spatial dimensions to 1×1 . This process has the dual advantage of drastically reducing the number of parameters while also minimizing overfitting. Subsequently, a fully connected layer is considered to enable learning non-linear combinations of features. The network ends with a softmax activation function, which

provides a probabilistic distribution of the output classes, enabling us to identify the most probable class for the input video.

5.4.2 Camera motion datasets

Due to the lack of available camera motion datasets, we have constructed our dataset from scratch. We consider the following seven distinct categories: static, pan left (PL), pan right (PR), zoom in (ZI), zoom out (ZO), tilt-up (TU), and tilt down (TD). For training and validation, we have followed a semi-automatic technique to annotate the videos, by leveraging traditional methods. For testing, we have acquired a dataset ourselves using different cameras.

5.4.2.1 Semi-automatic learning dataset creation

In a first step, we have collected a first video corpus which includes a number of 500 sequences, randomly gathered from YouTube [91]. We chose not to employ specific keywords like "camera", "pan", and "zoom" to refine our results because such videos are often shot with professional equipment and they tend to encapsulate cinematographic scenes. This could potentially inhibit the model's generalization ability.

The examination of the initial collection, shows that the majority of shots from these random videos are static. We did encounter some panning and tilting shots, but zooming shots were notably sparse. This led to a significant skew in our dataset. Consequently, we have specifically searched for "zoom" on YouTube, despite the potential lack of diversity, in an effort to increase the number of samples for the zoom class. We have also removed some of the static videos to improve the dataset balance.

The videos have been first segmented into shots, using the method introduced in [36]. This process has led to an approximate number of 3000 shots that needed to be categorized according to their corresponding camera motion type.

Figure 5.6 shows some examples of shots in this corpus, with various camera motion types.

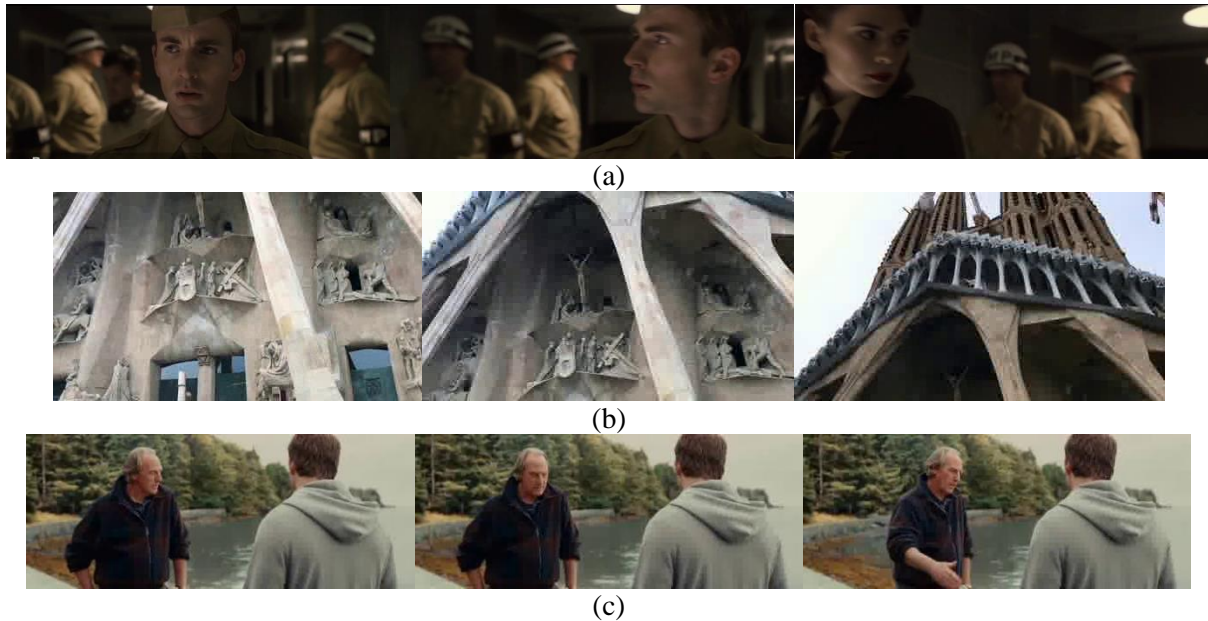


Figure 5.6. Examples from the videos of the dataset. (a) Pan Left, (b) Tilt Up, (c) Static

In order to perform the video shot categorization, we have adopted the method introduced in [41] and recalled in the following sections.

5.4.2.1.1 Interest point extraction

The first step in this annotation process involves the extraction of interest points from the videos. A regular grid sampling strategy is considered for this task, with a grid step size $\Gamma = \frac{W \cdot H}{N}$ (Figure 5.7). Here, W and H respectively represent the image's width and height, while N signifies the maximum number of points. The parameter N plays a critical role as it strikes a balance between processing time and detection accuracy. To keep computational costs relatively low while ensuring a high degree of precision, the value of N was set to 1000.



Figure 5.7. Grid of points in a frame.

5.4.2.1.2 Interest point tracking

The next step in the algorithm concerns tracking the considered interest points by calculating their displacement or motion vectors. To this purpose, the Lucas-Kanade algorithm [92] has been applied on each pair of consecutive frames in the video shot, starting from the initial one.

Let $p_{1i}(x_{1i}, y_{1i})$ be the i^{th} keypoint in the current image and $p_{2i}(x_{2i}, y_{2i})$ be its correspondent in the successive frame. The associated motion vectors (v_{ix}, v_{iy}) , expressed in polar coordinates with magnitude $(D_{i(1,2)})$ and angle of motion $(\theta_{i(1,2)})$ are expressed as follows:

$$v_{ix} = x_{2i} - x_{1i}; v_{iy} = y_{2i} - y_{1i} \quad (5.1)$$

$$D_{i(1,2)} = \sqrt{v_{ix}^2 + v_{iy}^2}, i = \overline{1, n} \quad (5.2)$$

$$\theta_{i(1,2)} = \arccos \frac{v_{ix}}{D_{i(1,2)}}, \theta \in [0, 2\pi] \quad (5.3)$$

where n is the total number of tracked points.

Figure 5.8 illustrates an example of tracked interest points between consecutive frames, for a panning motion type.

5.4.2.1.3 Background/ Foreground separation

The tracked interest points to determine the overall geometric transformation between two consecutive frames, represented as a homographic motion model. The RANSAC (*Random Sample Consensus*) [93] algorithm is here used to identify the optimal homographic matrix \mathbf{H} .



Figure 5.8. Correspondence between interest points in two successive frames.

Based on the matrix \mathbf{H} , for a given point $p_{1i}[x_{1i}, y_{1i}, 1]^T$ expressed in homogenous coordinates, its estimated position $p_{2i}^{est}[x_{2i}^{est}, y_{2i}^{est}, 1]^T$ is determined as:

$$\begin{bmatrix} x_{2i}^{est} \\ y_{2i}^{est} \\ w \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \cdot \begin{bmatrix} x_{1i} \\ y_{1i} \\ 1 \end{bmatrix} \quad (5.4)$$

where:

$$w = 1/(h_{20} \cdot x_{2i}^{est} + h_{21} \cdot y_{2i}^{est} + h_{22}) \quad (5.5)$$

The estimation error is defined as the difference between the estimated and the actual position of the considered interest point and is computed as:

$$e(p_{1i}, \mathbf{H}) = \|p_{2i}^{est} - p_{2i}\| \quad (5.6)$$

Key points with an estimation error $e(p_{1i}, \mathbf{H})$ below a threshold E (set to 2 pixels) are identified as part of the background (inliers). Conversely, those exceeding E are annotated as foreground objects (outliers).

5.4.2.1.4 Camera motion estimation

The estimation of camera motion relies on the inliers interest points tracked across frames. To ensure the reliability of the results, each frame is segmented into four distinct regions. The analysis is then performed independently on each regions and finally the resulting outcomes are compared in order to take the final decision.

Within each region, the dominant motion vector characteristics are determined by identifying the most common motion vector distances and angles (orientations). Let us denote by $D^{ur}, D^{ul}, D^{br}, D^{bl}$ (resp. $\theta^{ur}, \theta^{ul}, \theta^{br}, \theta^{bl}$) the prevalent distances (resp. angles) in the upper-right, upper-left, bottom-right, and bottom-left regions. The measurement of these values is then compared to predefined thresholds to discern the camera motion type between each pair of two consecutive frames. The specified thresholds are specified in Algorithm1.

ALGORITHM1: ALGORITHM TO ESTIMATE CAMERA MOTION

Input: Distance and angle estimation in the four regions, $D^{region}, \theta^{region} \forall region \in \{ur, ul, br, bl\}$

Output: Camera motion estimation \mathcal{F}^{CAM}

if $\forall region (D^{region} < 4 \text{ pix})$ **then**

$\mathcal{F}^{CAM} \leftarrow \text{Static}$

elif $\exists region (D^{region} < 4 \text{ pix})$ **then**

$\mathcal{F}^{CAM} \leftarrow \text{Unknown}$

elif $\exists region_1, region_2 (|D^{region_1} - D^{region_2}| > 20)$ **then**

$\mathcal{F}^{CAM} \leftarrow \text{Unknown}$

elif $\forall region (-180^\circ \leq \theta^{region} \leq -135^\circ \text{ or } 135^\circ \leq \theta^{region} \leq 180^\circ)$ **then**

$\mathcal{F}^{CAM} \leftarrow \text{Pan-Right}$

elif $\forall region (45^\circ \leq \theta^{region} \leq 135^\circ)$ **then**

$\mathcal{F}^{CAM} \leftarrow \text{Tilt-Down}$

elif $\forall region (-135^\circ \leq \theta^{region} \leq -45^\circ)$ **then**

$\mathcal{F}^{CAM} \leftarrow \text{Tilt-Up}$

elif $\forall region (-45^\circ \leq \theta^{region} \leq 45^\circ)$ **then**

$\mathcal{F}^{CAM} \leftarrow \text{Pan-Left}$

elif $\forall region (D^{region} > 1 \text{ pix})$ **then**

if $(90^\circ \leq \theta^{ur} \leq 180^\circ)$ and $(0^\circ \leq \theta^{ul} \leq 90^\circ)$ and $(-90^\circ \leq \theta^{br} \leq 0^\circ)$ and $(-180^\circ \leq \theta^{bl} \leq -90^\circ)$ **then**

$\mathcal{F}^{CAM} \leftarrow \text{Zoom-In}$

elif $(-90^\circ \leq \theta^{ur} \leq 0^\circ)$ and $(-180^\circ \leq \theta^{ul} \leq -90^\circ)$ and $(90^\circ \leq \theta^{br} \leq 180^\circ)$ and $(0^\circ \leq \theta^{bl} \leq 90^\circ)$ **then**

$\mathcal{F}^{CAM} \leftarrow \text{Zoom-Out}$

end

else

$\mathcal{F}^{CAM} \leftarrow \text{Unknown}$

end

The analysis has been performed for each video shot in the initial corpus. The shot is saved and stored in the final database if a uniform camera motion type is observed across a minimum number of 25 consecutive frames (which means that the same motion type is detected throughout all these frame successions).

Figure 5.9 presents some examples showcasing the dominant angle and distance across the four regions for eight classes: Static, Pan Left, Pan Right, Tilt-Up, Tilt-Down, Zoom In, Zoom Out, and Unknown.



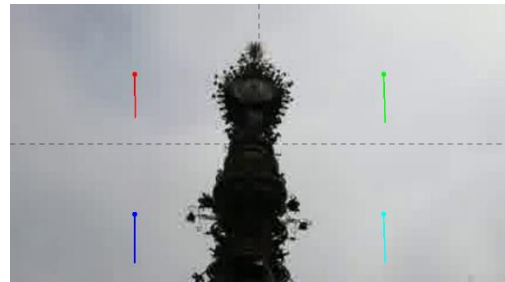
(a)



(b)



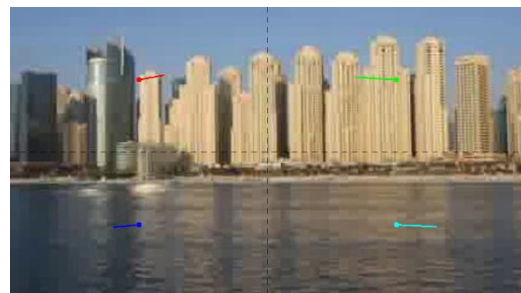
(c)



(d)



(e)



(f)



(g)



(h)

Figure 5.9. Dominant angle and distance across four regions in a frame. (a) Pan Right; (b) Pan Left; (c) Tilt-up; (d) Tilt Down; (e) Zoom In; (d) Zoom-out; (g) Static; and (h) Unknown.

5.4.2.1.5 Dataset cleaning

The considered method presents some limitations. Some failures have been encountered in close-up shots, where the foreground object occupy an important area of the scene, and blurry shots, where is difficult to reliably detect the interest points. In such cases, the video is classified as unknown and is not included in the dataset. In addition, the Lukas-Kanade tracking system cannot make the difference between static and slow-motion video segments. As a consequence, the process requires some human intervention (for less than 5% of the videos) to delete the outliers. However, let us underline that the goal of this method is solely related to the construction of a reliable training set that can serve to learn the 3D ResNet network, and which includes only correctly categorized video segments. The statistics of the obtained dataset are reported in Table 5.1. Some examples of the dataset are illustrated in Figure 5.10.

Table 5.1. The number of videos in each category in the train/val dataset

Category	S	PL	PR	TU	TD	ZI	ZO
No. of video segments	897	547	555	544	441	272	292



Pan Left



Pan Right



Static



Tilt Down



Tilt Up



Zoom In



Zoom Out

Figure 5.10. Examples from the Training dataset.

5.4.2.2 Creation of the test dataset

In order to conduct a comprehensive evaluation of our methodology, a test dataset has been specifically gathered under diverse acquisition conditions. This dataset not only encompassed videos of varying quality, including those with challenges such as blur, camera shake and changes in illumination or shadowing, but it also included a wide variety of scene types. This diversity in scene type ensured a robust representation of different shot types such as indoor and outdoor settings, close-ups and wide shots, as well as a variety of weather conditions, like sunny or rainy weather, and different times of day, covering both day and night scenarios.

The data collection has been undertaken using high-quality professional cameras, namely the Canon 5D Mark II and Sony Z280, supported with tripods for stability and made available by our partners at France Télévisions. We have also purposefully incorporated shots taken with smartphones into our dataset, as this allowed us to include video samples exhibiting hand-shake, a common factor in many real-world video capture scenarios. Furthermore, we considered a broad range of video resolutions and frame rates to ensure the adaptability and robustness of our method across different video quality and speed parameters. The resolutions varied from (1280×720) to (3840×2160) pixels, while frame rates ranged between 25 to 50 frames per second.

The videos included in our dataset generally last around 5 seconds each. These videos may contain several camera movements, which are interspersed with static sections. The distribution of video segments, defined as video intervals characterized by a singular camera motion, across the various categories is outlined in Table 5.2.

Table 5.2. The number of videos per category in the test dataset

Category	S	PL	PR	TU	TD	ZI	ZO
No. of video segments	450	369	422	261	263	251	252

Figure 5.11 illustrates some examples of video sequences from the test dataset.

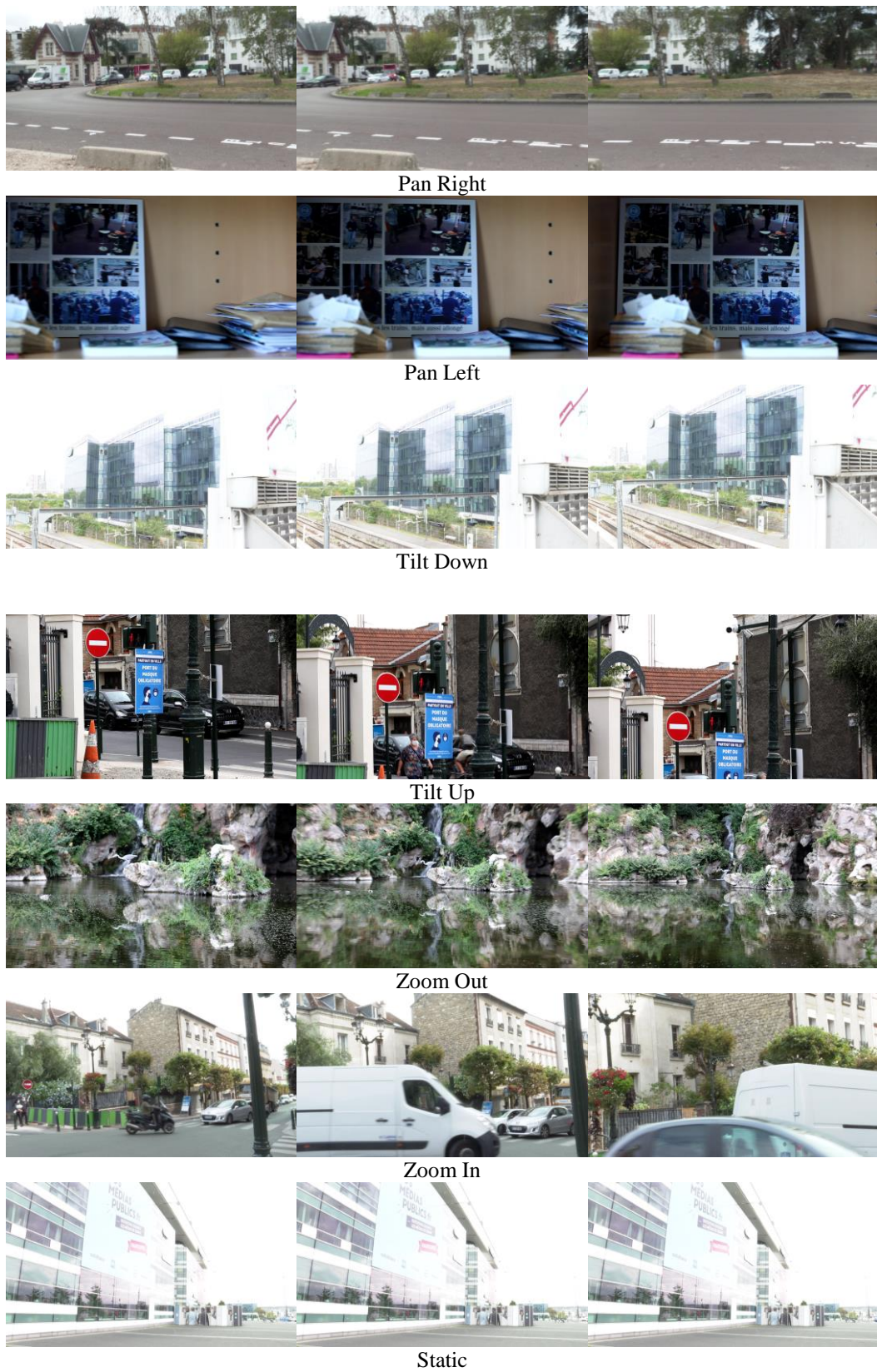


Figure 5.11. Examples from the test dataset. The videos includes high-resolution samples as well as hand-shake videos, blurry images and illumination variations.

5.5 Experimental results

In a preliminary stage, we have pre-trained the proposed network model on the generic action recognition Kinetics dataset [14]. The backpropagation algorithm minimizes the cross-entropy error, considered as loss function. For optimization, we have adopted a stochastic gradient descent with momentum. The weight decay was set to 0.001 and the momentum to 0.9. The learning rate started from 0.9 and was divided by 10 when the test loss saturated.

Then, the model has been retrained starting from the 3rd residual block with the training camera motion dataset extracted from Youtube videos (*cf.* Section 5.4.2.1). For fine-tuning, the learning rate has been set to 0.001 and the weight decay to $1e-5$. The learning process has been carried out on an Nvidia GTX 1080 GPU with batch size of 32. We randomly crop each video spatially and temporally to fit the network's input size $10 * 120 * 120 * 3$.

To increase scalability, we have also applied a data augmentation technique. To this purpose, we have reversed the order of the frames in each video and generate a new video sample with the opposite label (*e.g.*, if we reverse the frames of a zoom-in video we generate a new zoom-out video). This technique makes it possible not only to double the size of the dataset but also to increase the generality of the model.

Figure 5.12 shows the loss and accuracy curves obtained for four different conditions.

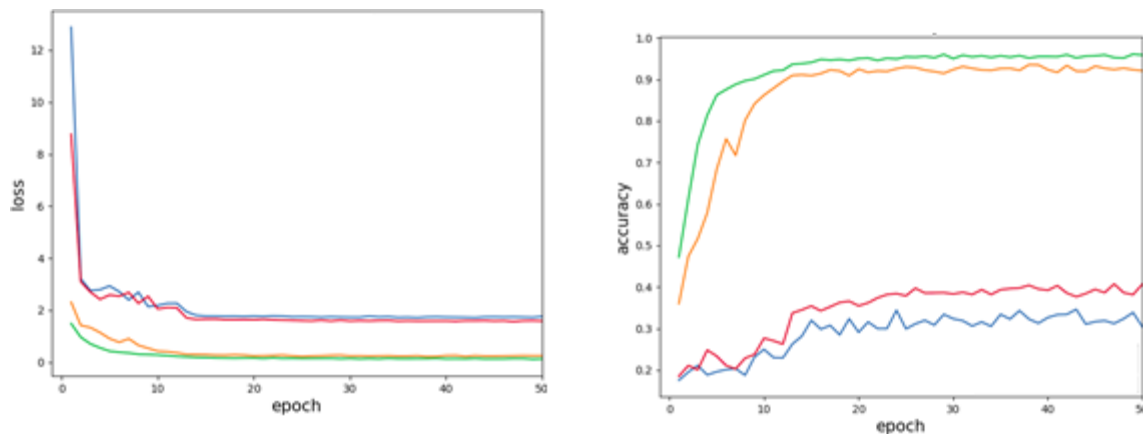


Figure 5.12. Comparison of the different configurations: (a). The loss variation (b). The accuracy variation. (Blue: Resnet trained from scratch, Red: Resnet + reverse frames, Orange: Resnet + finetuning, Green: Resnet + finetuning + reverse frames)

The baseline results are obtained by training Resnet-34 from scratch, without any preliminary training, and with our training dataset described in Section 5.4.1.

Without surprise, this approach resulted in poor results due to the low number of training samples. Augmenting the data by reversing the frames slightly improves the results but does not exceed an accuracy rate of 40%. These results confirm our intuition that such limited datasets cannot ensure a successful learning process. Let us now analyze the performances obtained for the transfer learning approach, where the 3D Resnet is first pre-trained on the Kinetics dataset and then fine-tuned on the augmented training set.

In the first time, we have examined the results obtained on the training data set derived from youtube videos. To this purpose, we have split the data into training and test sub-sets with a training/testing ratio of 80/20% and 5 cross-validation steps. In this case, the approach reaches 97% of accuracy.

Concerning the real-life test data set, since a given video may contain multiple camera motions, we have jointly performed segmentation and labeling. For this purpose, we have applied a sliding window technique with overlapping clips and stride 1, which means that the sliding window is successively slid by one frame. The size of the window is of intervals of 10 successive frames. At each frame position, the trained network estimates the class probabilities. As the video may contain several camera movements we use the static segments (a segment is a set of 2 clips or higher) to identify the end of a movement section. Then, we average the scores and the decision is taken based on a majority voting scheme over all frames of a movement section. The class with the highest probability is estimated as the correct camera motion.

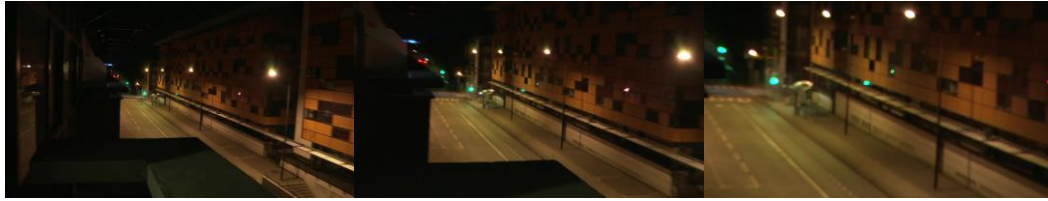
The resulting global, average accuracy rate obtained is 94%. These results, obtained on such challenging videos, fully demonstrate the pertinence of the proposed approach.

In a finer analysis, Figure 5.13 presents the confusion matrix obtained on the test dataset.

Static	0.97	0	0.01	0.01	0	0	0.01
PL	0.03	0.91	0.01	0	0.01	0.02	0.02
PR	0	0.02	0.94	0.02	0.01	0	0.01
TU	0.01	0	0	0.98	0	0.01	0
TD	0	0	0	0	0.98	0.02	0
ZI	0.02	0.03	0.01	0.01	0	0.93	0
ZD	0.01	0.02	0.01	0.01	0.02	0.02	0.9
	Static	PL	PR	TU	TD	ZI	ZD

Figure 5.13. Confusion matrix of the validation dataset

Figure 5.14 illustrates some results obtained on the test dataset. We can observe that some errors are occurring when the zooming is very slow. One explanation of this behavior may be the lack of diversity in the zoom class in the train dataset.



GT:ZI, P:ZI



GT:TD, P:TD



GT:PR, P:PR



GT:ZO, P:ZO



GT:TU, P:TU



GT:PL, P:PL



GT:ZO, P:S

Figure 5.14. Examples of recognition results on the test dataset.

5.6 Conclusion

In this chapter, we have introduced a data-driven camera motion classification method. The approach is based on a deep 3D Resnet model, under a transfer learning paradigm. In order to overcome the difficulties related to the poor availability of training/test camera motion data sets, we have proposed a transfer learning approach, which consist first in pre-training the model on a different purpose data set, related to action recognition, which is afterward fine-tuned on a lower scale, dedicated corpus. We have also introduced a semi-automatic technique for constructing such a corpus, starting from general-purpose videos. Finally, in order to test and validate our approach, we have acquired a validation data set, acquired in real-life conditions with different cameras and involving highly challenging videos. The experimental results obtained fully validate the proposed method, with an average accuracy rate of 94%. Our perspectives of future work mainly concern the extension of our approach with the help of active learning techniques that can further speed-up the training process and enhance the related performances.

PART II: MULTIMODAL MODELS

6 MULTIMODAL LEARNING

Abstract: The human perception is inherently multimodal, incorporating vision, hearing, touch, smell, and taste. A modality corresponds to the specific manner or channel through which data or experiences are captured or processed. Consequently, a research problem is classified as multimodal when it integrates several such modalities. In this chapter, we delve into multimodal learning, an expanding subfield of machine learning, which seeks to develop models capable of interpreting and learning from multimodal data. We review the key challenges in multimodal learning including data heterogeneity, fusion, alignment, and efficiency. Central to our discussion is an in-depth exploration of the transformer model, the latest state-of-the-art model for multimodal learning. We provide a detailed mathematical formulation of the vanilla transformer and we elaborate on the advantages and challenges of the model compared to previous architectures. Finally, we present the evolution and current state of multimodal datasets, which are fundamental to the development and benchmarking of new models.

Keywords: Multimodal learning, transformer.

6.1 Introduction

Multimodal learning (MML) is an area of machine learning that involves the integration and analysis of data from multiple types of data, such as text, images, audio and video, in order to perform predictions or derive meaningful conclusions. The primary significance of multimodal learning lies in its potential to encapsulate a richer context and a more comprehensive understanding of real-world scenarios than can be provided by unimodal learning, which operates on a single data type. In the context of human cognition, we naturally integrate multiple sensory inputs to perceive and interact with the world, hence, it is intuitive to apply the same principles to machine learning. This is why multimodal learning has emerged as a vital and active area of research. Its applications are diverse and far-reaching, ranging from sentiment analysis, where text and audio-visual cues can be combined to better understand users' sentiments, to autonomous driving, where various sensor inputs are integrated to safely navigate the vehicle. By harnessing multiple data modalities, we can create models that better mimic human intelligence and offer enhanced performance in complex, real-world tasks.

The initial part of this chapter is dedicated to a detailed discussion of the relevance of multimodal contexts within the domain of TV broadcasting, the primary application for our proposed methodologies. Subsequently, we turn our attention to the significant challenges encountered in the development of a multimodal deep learning model, covering critical considerations such as data heterogeneity, fusion, alignment and efficiency. We then delve into a thorough analysis of the prevailing multimodal architecture in contemporary literature - the transformer. This entails an examination of the mathematical underpinnings of the original transformer model, along with its advantages and challenges compared to previous methods. To conclude the chapter, we survey the current landscape of multimodal datasets, highlighting the intricate features they incorporate and how they have evolved over time.

6.2 Applications of multimodal learning for TV broadcast

Multimodal learning has found significant applications in the audiovisual broadcast community. Through the combination of text, video, and audio data, advanced algorithms can deliver more accurate video summarization and recommendation, enhancing viewer experience and providing personalized content that aligns with their preferences. Subtitle generation and translation, an essential aspect of international broadcasting, can also be significantly enhanced using multimodal learning, as context derived from audio and visual cues can provide more nuanced translations. Additionally, multimodal learning aids in advanced content analysis, enabling more precise categorization and tagging of shows or segments based on visual, auditory, and textual information. This could range from identifying key events in a sports broadcast to recognizing sentiment in a drama series. In the case of archive indexing, content such as JT (news), sports programs, and TV shows inherently comes in multiple modes. The indexing system, therefore, must be multimodal in order to effectively process, categorize, and archive such content.

In our research work, we have started by creating separate analyzers for different modes of data—video, image, and so forth, as detailed in Part I of the manuscript. Our objective in this part is to combine the insights gleaned from these separate analyzers, with the objective of constructing a more comprehensive and robust framework designed explicitly for applications considered by France TV.

We focus our development on two major axes. The first axis involves leveraging video question answering techniques. Video question answering systems are designed to respond to user queries about a specific video content. Applied to archive indexing, this technique allows users to ask specific questions about an archived content (examples: "Who appeared at the 15-minute mark of this news broadcast?" or "What was the score at halftime in this football match?") and retrieve precise answers.

The second axis revolves around the use of multimodal video captioning. Multimodal video captioning involves generating textual descriptions, expressed in natural language, using information from multiple modalities (typically audio and visual data). In the context of TV archive indexing, multimodal video captioning can facilitate a more accurate and nuanced understanding of archived content, supporting more precise search and retrieval capabilities. It goes beyond mere object identification or speech transcription. By combining visual, auditory, and possibly other sensory cues, the system can provide context-rich captions, effectively "summarizing" video content in a searchable text format. This could be incredibly useful for categorizing and retrieving relevant segments from archived footage based on viewer's or researcher's requirements.

Let us first present and analyze the multiple methodological challenges that need to be addressed when performing multimodal learning.

6.3 Challenges in multimodal learning

Multimodal learning offers a richer understanding of data compared to unimodal learning by simultaneously harnessing information from multiple modalities. However, this increased richness and complexity also presents various challenges. These challenges, ranging from data heterogeneity to the intricacies of fusion and alignment, necessitate innovative methodologies to fully exploit the potential of multimodal data for various applications. In this section, we delve deeper into these challenges and explore the strategies employed to tackle them in the context of multimodal learning.

6.3.1 Data heterogeneity

Data heterogeneity arises from the fundamental differences between modalities. Textual data is typically represented as sequences of words or relative embeddings [94], which capture semantic relationships and contextual information. In contrast, images are composed of grids of pixels or deep features derived from convolutional neural networks [19], emphasizing spatial information and visual patterns. Audio data is characterized by waveforms or spectrograms [95], representing temporal aspects and acoustic properties. There are various levels of heterogeneity in the context of multimodal data. For instance, when dealing with two languages that convey the same semantic meaning, the heterogeneity is lower compared to combining highly distinct modalities like textual data and sensor data [96]. Understanding the level of heterogeneity between modalities is pivotal when developing multimodal models. It guides decisions regarding encoder architecture, fusion methods, and the overall model design. For instance, in cases of low heterogeneity, a shared encoder [97]–[101] may be viable, reducing model complexity. In contrast, severe heterogeneity necessitates careful consideration of modality-specific encoders and sophisticated fusion mechanisms [102], [103].

6.3.2 Data fusion

The underlying integrative process, known as fusion, is a critical and multifaceted challenge in multimodal learning. It involves harmonizing information from different modalities to create a unified and coherent representation for decision-making or analysis. Fusion techniques determine how well a model can exploit the complementary information offered by different modalities. The goal is to leverage the strengths of each modality while mitigating their limitations. Effective fusion leads to improved performance, richer insights, and more accurate predictions in tasks such as video captioning and video question answering.

Multimodal fusion can occur at different levels of the processing chain. Several fusion strategies have been devised to tackle this challenge:

- *Early fusion*: This approach involves merging raw data or feature representations from multiple modalities at the input level of the model [104]. Early fusion aims to capture cross-modal relationships from the outset, creating a single, unified input for subsequent processing.
- *Late fusion*: Late fusion, conversely, maintains modality-specific processing until later stages of the model. Each modality is processed independently, and the results are combined at a higher level or during decision-making [105]. Late fusion provides flexibility in handling modality-specific characteristics.
- *Hybrid fusion*: Hybrid fusion combines elements of both early and late fusion to leverage the strengths of each approach. This strategy allows for multiple stages of fusion within a model, where early fusion might occur to capture certain cross-modal interactions, followed by late fusion to maintain the independence of modalities at later stages [106]. Hybrid fusion aims to strike a balance between preserving the unique information in each modality and exploiting cross-modal dependencies, making it suitable for complex tasks that involve a mix of correlated and distinct modalities.
- *Attention Mechanisms*: Attention mechanisms enable the model to dynamically weigh the importance of information from different modalities based on context or relevance to the task [107], [108]. These mechanisms have proven valuable in capturing salient features from each modality, particularly in tasks with varying levels of modality importance.

While fusion is central to multimodal learning, it is not without challenges. Selecting the appropriate fusion strategy depends on several factors, including the nature of the data, task requirements, and the degree of heterogeneity of the various modalities involved. Inadequate fusion strategies can hinder rather than enhance performance.

6.3.3 Alignment

In multimodal learning, alignment refers to the process of synchronizing or mapping data and information from different modalities in such a way that corresponding elements from each modality are related and matched appropriately [109]. The goal of alignment is to create a coherent and meaningful connection between the data sources, allowing for a better understanding, analysis, or joint processing of multimodal data. There are two fundamental approaches: explicit and implicit alignment. Explicit alignment refers to the precise synchronization or mapping of elements between different modalities. For instance, when aligning a video with its transcript, explicit alignment would involve matching specific video frames or segments with corresponding spoken words or subtitles, resulting in a clear and well-defined correspondence [110]. On the other hand, implicit alignment takes a more abstract or holistic approach, focusing on capturing underlying relationships and semantic connections between modalities without pinpointing exact correspondences [74]. Implicit alignment techniques often rely on advanced machine learning models, to learn complex patterns and associations between modalities, allowing for a more flexible and context-aware alignment that can adapt to variations and nuances in the data. Both explicit and implicit alignment have their unique strengths and applications, offering versatile tools for addressing alignment challenges in multimodal learning.

6.3.4 Efficiency

Multimodal learning is in general more resource-intensive and computationally complex when compared to unimodal learning. Efficiency in multimodal learning encompasses various dimensions, each posing distinct challenges:

- **Model size and parameters**: Multimodal models like CLIP [111] scale up significantly, surpassing 400 million parameters compared to unimodal models.

- Training data volume: Multimodal models require vast datasets, e.g., CLIP relies on 400GB of text and images, posing challenges in data collection and curation.
- Computational infrastructure (hardware): Multimodal models demand specialized hardware like GPUs or TPUs for training due to increased computational load [111].
- Inference latency: Real-time applications with multimodal models can experience higher inference latency, impacting user experience with response times ranging from hundreds of milliseconds to seconds [111].

As the demand for multimodal models continues to surge, the efficiency challenge takes a central place. This challenge becomes particularly critical with the advent of large-scale multimodal training, which has been popularized through the introduction of the transformer architecture and its adaptation to multimodal settings. In the following section, we introduce the transformer architecture, which has become *de facto* the most popular model for multimodal learning.

6.4 Transformer architecture for multimodal learning

In the field of MultiModal Learning (MML), the past few years have witnessed a transformative shift in the architectural landscape of models used. While monolithic models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been extensively employed, the arrival of transformer architectures [4] has sparked a major reorientation in research directions and techniques. These architectures have emerged as a beacon of promise due to their intrinsic advantages and immense scalability, offering a robust framework to model diverse modalities such as language, visual, auditory, and different tasks like language translation, image recognition, and speech recognition [100], [112]–[114].

Monolithic models, characterized by a single, consistent architecture, were traditionally tailored to work best with a specific type of data or task [115]. They underlie on modality-specific architectural assumptions, such as translation invariance in vision tasks for CNNs or sequence dependency in language tasks for RNNs. While such models have brought significant advancements to their respective domains, their ability to accommodate and learn from different modalities simultaneously remains limited.

On the other hand, transformer models stand out with their flexible and agnostic architecture [112]. Unlike their monolithic counterparts, transformers are designed to handle one or multiple sequences of tokens, irrespectively of the modality. This architecture, characterized by self-attention mechanisms, naturally lends itself to multimodal learning, without necessitating architectural modifications. Each sequence's attribute, such as the modality label or the sequential order, can be factored in, allowing transformers to comprehend per-modal specificity and inter-modal correlations effortlessly by merely controlling the input pattern of self-attention. Critically, this versatility and adaptability of transformer architectures have spurred a recent surge in research attempts across various disciplines. This has led to an explosive development of novel MML methods in recent years, paving the way for remarkable and diversified advancements in a multitude of areas.

The Transformer architecture, initially designed for natural language processing tasks, has since evolved into a versatile framework for multimodal learning. To grasp its functioning, let us begin by revisiting the mathematical formulation of the vanilla transformer.

6.4.1 Mathematical formulation of the Vanilla transformer

The pioneering Vanilla Transformer [4] architecture is fundamentally built on an encoder-decoder framework and utilizes tokenized input. It is composed of multiple layers or blocks of Transformers, as illustrated in Figure 6.1. These blocks incorporate two sub-components: a multi-head self-attention (MHSA) layer and a position-wise fully-connected feed-forward network (FFN).

6.4.1.1 Multi-Head Self Attention

At the heart of the Transformer lies the self-attention mechanism, which enables the model to weigh the importance of different elements within a sequence. Let us denote the input sequence as $X = (x_1, x_2, \dots, x_n)$, where n is the sequence length. The self-attention mechanism computes a new representation for each element x_i based on the entire input sequence X .

For each position i , the mechanism calculates an attention score $a_{i,j}$ with respect to every other position j in the sequence. These scores are determined through a compatibility function, typically implemented as a dot product:

$$a_{i,j} = \text{softmax}\left(\frac{Q(x_i) \cdot K(x_j)}{\sqrt{d_k}}\right) \quad (6.1)$$

Here, $Q(x_i)$ and $K(x_j)$ are linear projections of the input elements x_i and x_j using learned weight matrices. The scaling factor $\sqrt{d_k}$ helps stabilize the gradients during training. The softmax function ensures that the attention scores across all positions sum to 1.

With the attention scores established, the attention is computed as the weighted sum of values $V(x_j)$ at each position:

$$\text{Attention}(X) = \sum_{j=1}^n a_{i,j} \cdot V(x_j) \quad (6.2)$$

where the values $V(x_j)$ are linear projections of input elements x_j .

This mechanism allows the model to give more weight to relevant positions in the input sequence while suppressing irrelevant ones.

The Transformer architecture extends self-attention with multi-head attention, enabling the model to focus on different parts of the input sequence simultaneously. This is achieved by projecting the input into multiple subspaces, computing self-attention for each subspace, and then concatenating the results. Mathematically, for h attention heads, the multi-head attention operation can be defined as:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) \cdot W_o \quad (6.3)$$

Where head_i represents the output of the i -th attention head, and W_o is a learned weight matrix.

6.4.1.2 Position-wise Feed-Forward Networks

In addition to self-attention, the Transformer architecture includes position-wise Feed-Forward Networks (FFN) at each position in the sequence. These networks consist of fully connected layers with ReLU activations:

$$\text{FFN}(X) = \text{ReLU}(X \cdot W_1 + b_1)W_2 + b_2 \quad (6.4)$$

Here, W_1, b_1, W_2 and b_2 are learned parameters.

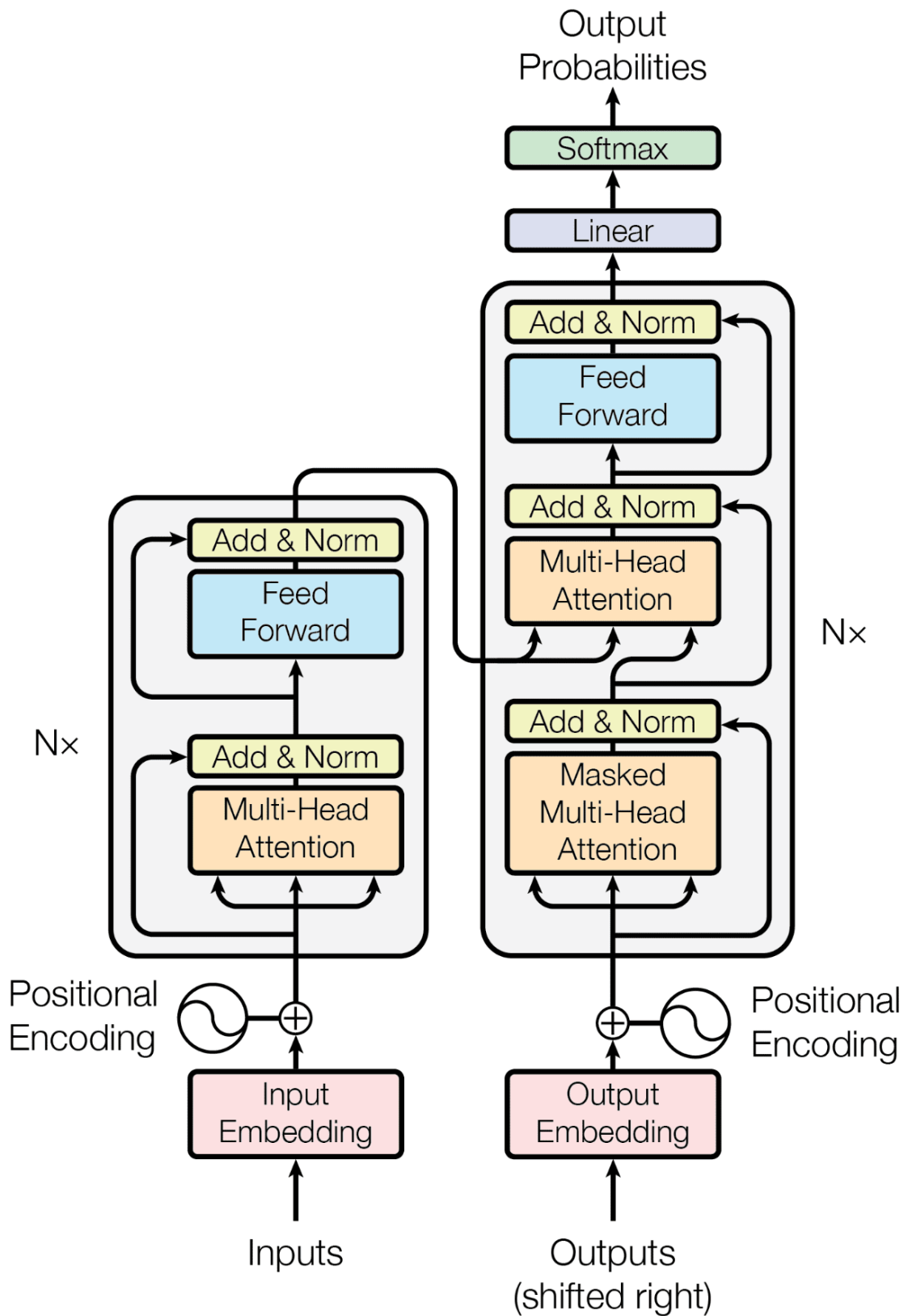


Figure 6.1. Architecture of the vanilla transformer [4].

6.4.2 Advantages of the transformer architecture

Let us now discuss the factors that have propelled Transformers to surpass traditional monolithic models like CNNs and RNNs in performance and capability.

Parallelization: Transformers can process sequences in parallel rather than sequentially, making them highly efficient for tasks that involve long sequences. This parallelization leads to faster training and inference times compared to RNNs, which process sequences one element at a time.

Versatility: Monolithic models have fixed architectures optimized for specific modalities (e.g., images for CNNs, sequences for RNNs). Adapting them to handle multiple modalities requires significant architectural modifications. On the other hand, transformers are composed of layers and attention heads, providing a modular structure that can be adapted and extended for various tasks. This modularity simplifies architecture design and experimentation.

Cross-modal interaction: Transformers employ attention mechanisms that inherently enable cross-modal interaction. This is in contrast to CNNs, where incorporating cross-modal interaction often requires complex fusion layers and handcrafted architectures. Transformers' attention mechanisms allow them to naturally attend to and capture cross-modal dependencies, making it more straightforward to model interactions between different modalities.

Long-range dependencies: Transformers excel in handling long-range dependencies compared to traditional monolithic models due to their fundamentally different architecture. The attention mechanism allows every element in a sequence to consider and weigh the importance of all other elements, regardless of their distance. As a result, transformers can efficiently capture both short-range and long-range dependencies in data, making them highly effective for tasks requiring extensive contextual understanding. This ability to maintain consistent gradients, parallelize computations, and incorporate positional encoding ensures that transformers can model complex relationships across extended sequences with ease, setting them apart as the preferred choice for tasks demanding the modeling of intricate long-range dependencies.

Large-scale pre-training: While monolithic models have been adapted and extended for multimodal tasks, the transformative capabilities of unsupervised large-scale pre-training, as seen in transformers, were not initially present. Transformers can be scaled to handle massive amounts of data and large models with billions of parameters. When fine-tuned on specific downstream tasks, the pre-trained transformers consistently deliver state-of-the-art performance. This phenomenon has catalyzed the democratization of unsupervised large-scale pre-training, effectively liberating practitioners from the labor-intensive process of data labeling. By offering a powerful foundation of pre-learned features, these models reduce the manual burden, accelerate experimentation, and enable a more inclusive approach to advanced machine learning.

6.4.3 Challenges in transformers

The key challenges that have emerged with the widespread adoption of transformer models are the following.

Fixed sequence length: Unlike RNNs, which can naturally handle variable-length sequences, transformers require input sequences of fixed length, limiting their applicability in tasks with varying context lengths. To overcome this limitation, researchers have proposed techniques such as segmenting long sequences into smaller chunks, employing hierarchical models, or padding shorter sequences with a special "[PAD]" token to match the length of the longest sequence in a batch [116]. These approaches aim to make transformers more flexible in handling data with diverse and dynamic context lengths, widening their scope of applicability beyond their original fixed-length constrain

Performance in data-constrained environment: Transformers, renowned for their exceptional performance in data-rich environments, can face challenges in data-constrained settings. Their reliance on large-scale pre-training and massive corpora can be a limitation when labeled data is scarce. In such scenarios, monolithic models that incorporate strong inductive biases may outperform transformers [117]. Monolithic models often have built-in structural assumptions tailored to specific modalities or tasks. These biases can help them excel with limited data, as they inherently encode prior knowledge about the data domain. In contrast, transformers rely more on learned representations and may struggle when the training data is sparse.

Interpretability: Interpretability has been a recurring challenge in transformer models due to their complex architecture, particularly the self-attention mechanisms [118]. Understanding why a transformer makes a specific prediction or which parts of the input data it focuses on can be elusive. Addressing the interpretability problem involves various strategies. One approach is attention visualization [4], [108], where researchers visualize the attention weights of the model to gain insights into its decision-making process. Attention maps can help identify which tokens in the input sequence are most influential in making predictions. Additionally, techniques like attention probing and feature attribution methods aim to dissect the internal workings of transformers [119]–[121]. These methods can shed light on what information the model finds important for specific tasks. Ongoing research focus on making transformers more interpretable, which is crucial for their broader adoption in critical applications where model transparency is essential.

Efficiency: The efficiency challenge in transformers encompasses several critical aspects that have garnered substantial attention from researchers. Transformers exhibit a quadratic increase in computational complexity with the length of input sequences, posing challenges in processing long sequences efficiently. Their resource requirements, including memory and high-capacity hardware like GPUs and TPUs, can strain computational infrastructure. This, in turn, contributes to higher computational complexity, slowing down both training and inference. In the context of multimodal learning, these efficiency concerns become even more crucial, given the inherent complexity of combining and processing diverse data modalities. State-of-the-art solutions have emerged to address such challenges, including knowledge distillation [122] to transfer knowledge from large models to smaller ones, model compression and pruning techniques [123], and sparse factorization of the attention matrix [124] to reduce computational demands. These innovations are pivotal in making transformer-based models more practical and accessible in resource-constrained environments, facilitating their broader application across a wide range of domains.

The transformative influence of transformer models on deep-learning methodologies cannot be overstated. Their resounding success is primarily attributed to their unique architecture that excels at capturing long-range dependencies and complex patterns in data through attention mechanism. This permits sophisticated understanding of contextual relationships in diverse data sets, thereby significantly boosting model performance [117]. In this context, the research community has invested considerable effort to enhance the scale and diversity of datasets to allow for better generalization of models. In the next section, we review the current landscape of multimodal datasets.

6.5 Landscape of multimodal datasets.

In the past decade, the boom in internet applications, particularly social media and online retail, has led to an explosion in the generation of vast multimodal datasets. Examples of these datasets are numerous and span a variety of contexts, from Conceptual Captions [125] and COCO [126] to VQA [127] and Visual Genome [128], among many others.

Recently, we have observed the emergence of several new trends in these datasets that reflect the evolving landscape of multimodal machine learning:

Larger data scales: In response to the demands of increasingly complex machine learning models, we are witnessing the creation of datasets on an unprecedented scale. These include datasets such as Product1M [128] and Conceptual 12M [125] that comprise millions of data points. These large-scale datasets offer a wealth of information for training more robust and nuanced machine learning models such as transformers.

More modalities: Beyond the traditional vision, text, and audio modalities, new diverse ones are being incorporated to reflect a broader spectrum of human experience and perception. For instance, Pano-AVQA [129] and YT-360 [130] incorporate 360° videos, while AIST++ [131] combines 3D dance motion with music. These diverse modalities push the boundaries of current machine learning capabilities and offer exciting opportunities for future research and applications.

More scenarios: Datasets are being tailored to explore a wider array of application contexts, moving beyond generic scenarios to more specific and niche applications. Examples of this include the use of real-life images in CIRRR [132], financial data in M3A [133], and autonomous driving data in X-World [134]. Such highly specialized datasets facilitate the development of models with more targeted and nuanced capabilities.

More complex tasks: To challenge and push the capabilities of current machine learning models, datasets are featuring more abstract and complex multimodal tasks. These include metaphor understanding in the MultiMET dataset [135] and hate speech detection in the Hateful Memes dataset [136]. These tasks compel models to delve into more complex forms of understanding and inference that are closer to human cognition.

Instructional videos: There is an increasing popularity of instructional videos, such as YouCookII [110] and HowTo100M [137] where machine learning models are tasked with aligning a sequence of instructions with someone performing a task in a video [138]. This serves as a powerful pre-training task [100], as it forces models to understand temporal dynamics and causality in real-world actions, mirroring how humans learn from instructions and demonstrations.

As with other deep learning architectures, Transformers are data-hungry. Their effectiveness in multimodal machine learning is due in part to the symbiosis between their high-capacity models and the availability of extensive multimodal data. This relationship has even enabled zero-shot learning capabilities in certain VLP (Video-Language Pre-training) Transformer models, demonstrating the potential of such models when trained with rich and diverse multimodal data.

6.6 Conclusion

This chapter provided a comprehensive examination of multimodal learning with a particular focus on deep-learning-based approaches. Central to the discussion was a detailed exploration of the primary challenges in multimodal learning, including data heterogeneity, fusion, alignment and efficiency. Our review highlighted the revolutionary role the transformer architecture has played in multimodal learning, firmly establishing itself as a fundamental network over the past five years. We presented the mathematical formulation of the vanilla transformer, providing an in-depth understanding of its theoretical framework. The latter part of the chapter extended the discussion towards the key advantages and difficulties when applying the transformer architecture to multimodal learning. Concluding this chapter, we presented the benchmark datasets in multimodal learning and the evolving features they encapsulate.

In the next chapters, we will delve into two main tasks in multimodal learning: Video Question Answering and Video Captioning. We chose the video channel as it represents a rich source of information, fusing visual and auditory cues over time. This multimodal nature makes it an excellent

testing ground for advanced MML techniques. Throughout the chapters, our primary focus will be on visual-textual training, elucidating its pivotal role in bridging the boundaries between computer vision and natural language processing methodologies. We will also discuss the application of the transformer model as a universal architecture for the stated tasks.

7 VIDEO QUESTION ANSWERING

Abstract: Video question answering (VideoQA) is the process that aims at providing a semantically pertinent answer to questions expressed in natural language, related to the content of a given video. VideoQA is a highly challenging task and requires a comprehensive understanding of the video document, including the recognition of the various objects, actions and activities involved together with the spatio-temporal relations between them. In this chapter, we introduce a novel VideoQA method, based on a conditional cross-correlation network that is able to learn a multimodal contextualization with reduced computational and memory requirements. At the core of our approach, we consider a cross-correlation module designed to learn reciprocally constrained visual and textual features combined with a lightweight transformer that fuses the intermodal contextualization between the two modalities. In addition, a video transformer with temporal attention is introduced to learn contextual features from the video. We also automatically extract the transcript of the video, which is considered as an additional modality, and investigate its impact on the model's performance. The vulnerability of the composing elements of our framework is tested using black box attacks that represent automatically-generated, semantic-preserving rephrased questions. The experimental evaluation, carried out on the MSVD-QA and MSRVTT-QA benchmark datasets, validates the proposed methodology with average accuracy scores of 44.96% and 41.88% respectively. When compared with state-of-the-art methods the proposed method yields gains in accuracy of more than 2%.

Keywords: Video Question Answering, multimodal learning, cross correlation.

7.1 Introduction

Video Question Answering (VideoQA) involves predicting an accurate answer a^* based on a question q and a corresponding video V (Figure 7.1). There are primarily two categories of tasks within VideoQA: multi-choice QA and open-ended QA. In the case of **Multi-Choice QA (MC)**, models are given several potential answers A_{mc} per question, with the objective of identifying the correct one $a^* = F(a|q, V, A_{mc})$. **Open-Ended QA (OE)**, on the other hand, can take the form of classification (most common), generation (word-by-word), or regression (typically used for counting), each depending on the specific datasets. More frequently, open-ended QA is defined as a multi-class classification problem, requiring the models to assign a video-question pair to a predefined answer vocabulary set A_{oe} as follows: $a^* = F(a|q, V)$ where $a \in A_{oe}$. This task can also be formulated as a generation problem, which is gaining increasing attention due to its practical utility. Generally, the answer is represented as a vector $a = (a_1, a_2, \dots, a_t, \dots, a_M)$ of length M . The prediction of the t^{th} word a_t^* is formulated as $a_t^* = F(a_t|q, V, (a_1, \dots, a_{t-1}))$.

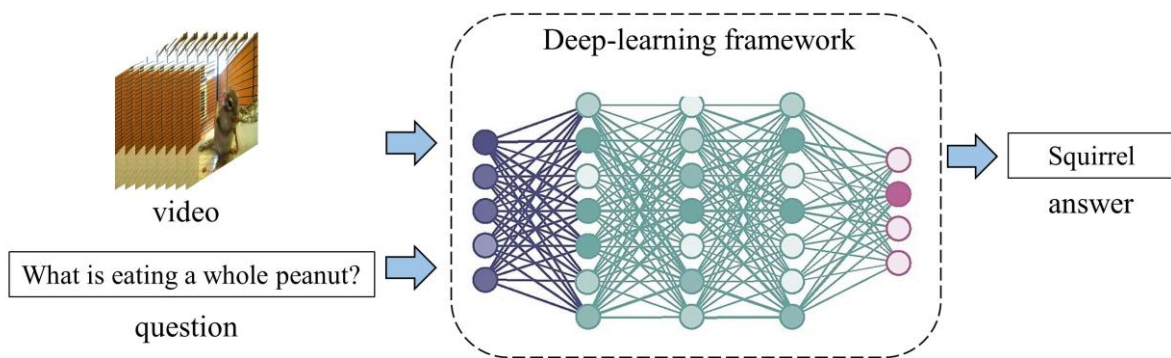


Figure 7.1. Video Question Answering task.

The remainder of the chapter is organized as follows.

In Section 7.2, we delve into the industrial application of video question answering, particularly reviewing its usage for the indexing and retrieval processes in the archives of France TV.

Section 7.3 presents a detailed review of the current state of the art in the VideoQA task, including a discussion on benchmark datasets, related evaluation metrics, and various families of methodologies involved. We categorize these methodologies into monolithic models with attention, memory-based models, graph-based models, and transformer-based models, shedding light on the advantages and disadvantages of each approach.

The analysis of the literature, with limitation of existing methods conducted us to propose our own approach, introduced in Section 7.4. We first detail the extraction of features from both visual and textual modalities. We then explore the correlation between these features using a cross-modal module designed to address the alignment issue, prevalent in multimodal learning. Lastly, we implement a transformer encoder to learn grounded visual-textual features through attention mechanisms, and predict the answer using a classification head over the answer vocabulary.

In Section 7.5, we evaluate the robustness of our model against adversarial attacks, an important measure to assess the model's generalization to unseen data. This starts with an overview of the problem and current methodologies, followed by the explanation of our methodology for implementing rephrasing attacks on the model.

Section 7.6 provides an in-depth experimental evaluation on the considered datasets: MSRVTT-QA and MSVD-QA. We conduct extensive ablation studies to evaluate the significance of each building block of our framework, and subsequently compare our approach to previous state-of-the-art methods.

Finally, in Section 7.7 we propose potential research areas that warrant further exploration in the future.

7.2 Application of Video Question Answering for archive indexing and retrieval

The application of video question answering techniques holds significant potential for enhancing archive indexing and retrieval processes. One of the challenges that documentalists face during the indexing process, is related to the subjective nature of determining the importance of events within a video. The relevance of certain information can vary from one individual to another based on their understanding and perspective. For instance, a documentalist with a keen interest in politics may highlight a subtle policy implication in a news broadcast that others might overlook. This subjectivity is further complicated by the temporal dimension. Events or details deemed inconsequential today may gain importance in the future due to evolving social, political, or cultural contexts. For instance, a casual comment about climate change in a decade-old newscast may assume greater significance today under the light of recent climate crises.

One potential solution to this challenge is exhaustive indexing, capturing every bit of information within a TV program. However, this approach tends to generate an overwhelming amount of data, often including irrelevant information, making the retrieval process cumbersome and inefficient.

An alternate and more promising solution is the application of video question answering techniques. In this setting, videos are encoded into compact feature vectors that encapsulate a comprehensive understanding of the video content. These vectors are derived from multimodal information—audio, visual, and possibly textual cues—allowing a depth of understanding beyond mere surface-level description. For instance, in a news broadcast about a political rally, a video question answering system could capture not just the who, what, and where, but also the sentiment, the context, and the implications.

During the retrieval process, real-time queries can be rapidly searched within the dataset. For example, a user could ask, "Which broadcasts feature discussions on the impact of the recent tax reform?" or "Show me segments where the president addresses climate change." The system, having a rich understanding of the content, can then deliver precise, relevant results, making it an efficient tool for navigating vast video archives. Thus, video question answering techniques promise to revolutionize archive indexing and retrieval by overcoming subjectivity and temporal shifts in importance, thereby greatly enhancing the accessibility and usability of archival footage.

7.3 Related work

Let us first review the various datasets and evaluation metrics largely employed today for training VideoQA models.

7.3.1 VideoQA datasets and evaluations metrics

7.3.1.1 VideoQA datasets

VideoQA datasets can be examined through various views, offering an array of perspectives from which these datasets may be assessed and categorized. One significant way to classify these datasets is based on the modalities used (visual-based, multimodal or knowledge-based). Another key dimension for classification concerns the complexity of the questions asked in the dataset (factoid or inference). Additional classifications can be made based on factors such as the length of the video (long-form or short videos) or type of the video (natural videos, GIFs, synthetic videos), and so forth.

In this section, we address two key taxonomies: Modality-based classification and Question-based classification. These areas of focus have been selected due to their fundamental role in shaping the structure, functionality, and evaluation of VideoQA datasets. In Table 7.1, we summarize the key statistics of video question answering datasets.

7.3.1.1.1 Modality-based classification

Video Question Answering (VideoQA) can be classified into distinct categories based on the data modality. These categories include visual-based VideoQA, multi-modal VideoQA (MM VideoQA), and knowledge-based VideoQA (KB VideoQA). It is important to note that the VideoQA task is inherently multimodal, as the model is tasked with processing both video and question inputs. Through the application of multimodal and visual-based classification, our objective is to differentiate between the modalities engaged in video analysis - whether exclusively visual cues are used, or if additional modalities are incorporated. Figure 7.2 presents some examples from various datasets.

Visual-based (VB) datasets [1], [139], [140] are characterized by the sole reliance on visual cues to comprehend the question and deduce the accurate answer. This category underscores the visual comprehension of video components and the reasoning of their interrelationships. In general, the videos employed in this category tend to be short and user-generated, notably from social platforms.


Multi-Modal VideoQA [141]–[143] frequently integrates resources beyond visual content, such as subtitles or transcripts and textual plots of movies [142] and television shows [143]. The principal challenges posed by MM VideoQA revolve around the fusion of multi-modal information and the comprehension of extensive video narratives.

Finally, **knowledge-based** (KB) VideoQA [NO_PRINTED_FORM] necessitates the distillation of external knowledge from explicit knowledge bases or the application of commonsense reasoning [144]. KB VideoQA requires a global knowledge base for the entire dataset, as opposed to supplying paired "knowledge" for each individual question.

7.3.1.1.2 Question-based classification

VideoQA datasets can be categorized into two distinct types based on the nature of the question or the challenges posed within the questions: factoid VideoQA and inference VideoQA. **Factoid VideoQA** (FQA) [1], [139], [141] involves questions that directly seek visual facts, such as queries about location ("where is"), objects or attributes ("who/what (color) is"). These questions require minimal relational understanding to interpret the questions and generate the correct answers. Emphasis in Factoid QA is placed on a comprehensive understanding of the questions and the identification of the visual elements involved.

On the other hand, **Inference VideoQA** (IQA) [145]–[147] is designed to examine the capacity for logical reasoning and knowledge application within dynamic scenarios. It features a variety of relationships between the visual facts such as temporal ("before/after") and causal ("why/how/what if") relationships.




Well, this is the mother of all earthquake faults. It can pack wallop 30 times that of the San Andreas fault. So forget all the Hollywood hype about the San Andreas fault. We're talking about an earthquake a night.

What is a man talking on tv about?

Earthquake

(a)



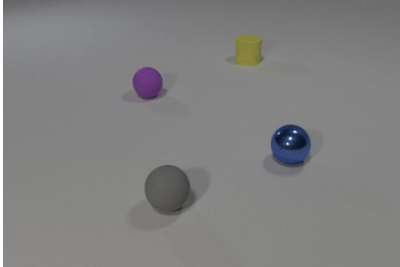
Howard (jumping off game mat): Grab a napkin, homey, you just got served.
Leonard: It's fine. You win. Howard: What's his problem?
Sheldon: His imaginary girlfriend broke up with him.

What girlfriend is Sheldon talking about?

Oracle: Penny was angry at Leonard in this episode.

a) Priya b) Amy c) Bernadette d) Penny

(c)



What is the shape of the object to collide with the purple object?

Sphere


(e)



How many people are riding camels?

Two

(b)




You said, green. Survey said.

Why is the woman in the red dress cheering?

a) She cheers because she answers a question correctly.
 b) She is very glad.
 c) She is upset that she lost.
 d) She is upset that she won.

(d)



What is eating a whole peanut?

Squirrel

(f)

Figure 7.2. Examples from different datasets. (a) MSRVTT-QA; (b) ActivityNet-QA; (c) KnowIT; (d) SocialIQ; (e) CLEVER; (f) MSVD-QA.

Table 7.1. Statistics of VideoQA datasets.

Dataset	VB	MM	KB	FQA	IQA	QA form	Video length (s)	#Clips	#QA pairs
MovieQA [142]		✓		✓		MC	200	6,771	14,944
MSVD-QA [1]	✓			✓		OE	10	1,970	50,505
MSRVTT-QA [1]		✓		✓		OE	15	1,970	50,505
TGIF-QA [140]	✓				✓	MC	3	56,720	103,919
MovieFIB [148]		✓		✓		OE	-	128,085	348,998
TVQA [143]		✓		✓		MC	76	21,793	152,545
ActivityNet-QA [139]	✓			✓		OE	180	5,800	58,000
Social-IQ [147]		✓			✓	MC	-	1,250	7,500
CLEVER [149]	✓				✓	MC/OE	5	20,000	305,280
KnowIT VQA [150]			✓		✓	MC	20	12,264	24,282
HowToVQA [141]		✓		✓		OE	12.1	69,270,581	69,270,581
iVQA [141]						MC	18.6	10,000	10,000

7.3.1.2 Evaluation metrics

Evaluating the semantic and syntactic correctness of computer-generated sentences is a notoriously challenging task because of their inherent ambiguity. There are several methods for evaluating text-generated systems such as BLEU [151], ROUGE [152], and METEOR [153], all of which measure the word overlap between ground truth and prediction. However, these values do not correlate well with human judgement and exhibit well-known blind spots [154]. To overcome this problem, most VideoQA datasets limit the number of their response domains to single words or short sentences.

Yu *et al.* [139] proposes two evaluation metrics to measure the performance of the models on their dataset.

Accuracy: Commonly used for evaluating classification tasks, accuracy is simply the ratio of the correct predictions to the total number of input samples. It is defined as:

$$Accuracy = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[a_i = y_i] \quad (7.1)$$

where a_i and y_i are the predicted and ground truth answers respectively and $\mathbb{1}[\cdot]$ is an indicator function that is equal to one only if a_i and y_i are identical.

Yang *et al.* [141] collects for each question at least 2 ground truth answers from different human annotators and compare the predicted answer with the human-generated ones. They define the accuracy as:

$$Accuracy(a_i) = \min\left(\frac{\#\text{ground truth answers} = a}{2}, 1\right) \quad (7.2)$$

The score assigns 1 if the predicted answer is confirmed by at least 2 annotators, 0.5 if it is confirmed by only one annotator and 0 otherwise. This accuracy is a reasonable metric for multi-choice datasets, but fails for open ended questions. For example, if the ground truth is “light green” the accuracy assigns 0 to the answer “green”.

WUPS: The Wu-Palmer similarity (WUPS) [155] accounts for word-level ambiguity and measures the similarity between the ground truth and the candidate answer by finding the longest common subsequence in the taxonomy tree. It is based on the WUP measure and considers WordNet [156] to calculate the distance in the semantic tree of words w and v contained in the predicted answer and ground truth, respectively.

$$WUPS = \frac{1}{N} \sum_{i=1}^N \left\{ \prod_{w \in a_i} \max_{v \in y_i} \mu_\gamma(w, v), \prod_{v \in y_i} \max_{w \in a_i} \mu_\gamma(w, v) \right\} \quad (7.3)$$

$$\mu_\gamma(w, v) = \{WUP(w, v) \text{ if } WUP(w, v) \geq \gamma \text{ } 0.1 \times WUP(w, v) \text{ otherwise} \quad (7.4)$$

The predicted answer is considered as correct only if the similarity between two words exceeds a pre-defined threshold. Similarly to [157], the metric is evaluated against two thresholds $\gamma = 0.0$ and $\gamma = 0.9$ and (called WUPS@0.0 and WUPS@0.9 respectively).

7.3.1.3 Discussion

The choices made by the dataset creators affect the complexity of the model to develop in many ways and may lead to some limitations.

As a main drawback let us first mention the limited size of the available datasets (Table 7.1). Without sufficient samples, the model under-fits during training, and test results unreliably reflect real-world performance. The problem is related to the techniques used to generate the question-answer pairs. On the one hand, human annotation is expensive, tedious, and difficult to scale. However, it provides more variety and level of abstraction. On the other hand, automatic techniques can generate larger datasets, but have significant limitations. Namely, the transcript data contains little information about the video, which leads to linguistic bias in training. In addition, the QA pairs collected are often redundant and lack variety, making the model subject to overfit. This technique might also generate incorrect predictions. Automatic conversion of descriptions into free-form question-answer pairs is still an open research topic. The lack of large datasets with trustworthy labels weakens the adoption and successful use of VideoQA systems in real applications.

Second, most benchmark datasets focus on short video content to facilitate semantic representation modeling (Table 7.1). However, the length of the video strongly correlates with the complexity of the answer prediction. The long-form videos often represent complex temporal interactions and causality through frames. Although models trained on short videos show promising performance, they may be ineffectively applied to long videos (e.g, TV shows and movies) due to the lack of representation of long-term semantic dependencies, making it difficult to distinguish the performance of different VideoQA approaches.

Third, most datasets address factoid questions such as "what," "who," and "where." These questions are attractive because they can be answered with a single word or short phrase, which facilitates system evaluation. However, they require a low level of computer vision. Recently, there has been a new interest in promoting causal and counterfactual questions ("how," "why," "what if ") [144], [147], [149]. Such

questions are asked by users to identify reasons and explanations about certain events or objects and require challenging VQA systems. The same question can have different interpretations (multiple answers) and require answers ranging from one sentence to a whole paragraph. To facilitate the development of such models, current datasets provide answers in a multi-choice environment. It is important to note that there is an ongoing discussion about the validity of this setting. For example, it is difficult to confirm the performance of a model given that it may be correct by chance. In addition, we cannot use such a model in real life, where there is no possibility of multi-choice responses. Moreover, there is a strong correlation between the question formulation and the correct answer obtained. That is, if we randomly generate the distractors, it would be easy to "guess" the correct answer based on the semantic similarity between the question and the answer.

Finally, existing datasets limit the number of words in the answers, by design, to facilitate evaluation of the model using simple metrics such as accuracy and WUPS. These metrics are useful for generating short sentences, but have serious drawbacks when it comes to comparing real-life responses. As research progresses, we expect to generate longer and more comprehensive sentences. For example, in Social-IQ [147] the average length of the answers is 10.46 words which is close to the average length of the captions in MS-COCO [158] (10.5 words). This is because the advanced questions lead to higher level concepts, such as object relationships, actions, etc., as opposed to factoid questions which only require a named entity (2 words). This should stimulate research to develop human-like scoring systems that can accurately assess the performance of VideoQA models.

In our work, we have retained two well-known benchmarks, namely MSVD-QA and MSRVTQ [1]. The MSVD-QA dataset consists of visual-based data with open-ended factoid questions, while the MSRVTQ dataset is a multimodal dataset containing videos and subtitles. We specifically opted for these datasets because research in these areas is still evolving and has not yet reached a mature stage. Notably, state-of-the-art methods achieve accuracy levels ranging from 40% to 45% on these benchmarks, leaving ample room for improvement. We aimed to explore these areas further using our approach. Additionally, it is worth mentioning that most knowledge-based and inference-based datasets are presented in a multiple-choice format, which does not effectively evaluate the model's performance and lacks real-life applicability.

7.3.2 State of the art VQA techniques

Video question answering systems are models that infer the correct answer to a natural language question from the content of a video. The *de facto* paradigm to solve this problem is to extract visual features using pre-trained vision models, and textual features using pre-trained language models, and then merge these representations into a common embedding space using a multimodal fusion module (Figure 7.3). The encoders for video and text can either belong to the monolithic models group (CNNs and RNNs) or the transformer family, such as BERT [116], ViViT [159], or ViT [112].

Within this context, we have identified three primary families of methodologies: Monolithic models (section 7.3.2.1), Memory-based approaches (section 7.3.2.2), Graph-based models (section 7.3.2.3), and Transformer-based approaches (section 7.3.2.4).

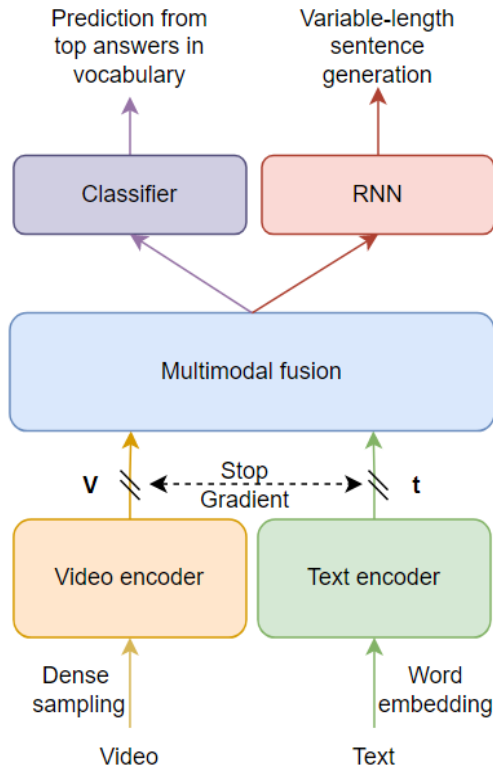


Figure 7.3. Basic VideoQA Framework.

7.3.2.1 Monolithic models with attention

In [160], Zeng *et al.* attempted to fuse global video and question representations for answer prediction using element-wise multiplication directly. The study highlighted the effectiveness of a straightforward temporal attention mechanism. This concept of attention was further investigated in more intricate scenarios, and was integrated with a variety of concepts, such as the multi-granularity ensemble [1] and hierarchical learning [161].

Specifically, Jang *et al.* [140] proposed a method based on dual-LSTM, employing both spatial and temporal attention mechanisms. Their approach showed the effectiveness of frame-guided attention and region-guided attention (using regions of interest ROI). Xu *et al.* [1] enhanced the attention mechanism over frame-level and clip-level visual features, guided by both the coarse-grained question feature and fine-grained word feature. Zhao *et al.* [162] introduced hierarchical dual-level attention networks (DLAN) designed to represent question-aware video representations, utilizing both word-level and question-level attention applied to appearance and motion.

However, it is important to note that despite the capability of these approaches to attend to video frames and clips, they depend on RNNs for history information modeling. This approach has subsequently been found to be relatively inefficient in capturing long-term dependencies [163].

7.3.2.2 Memory-based models

Memory-based networks have the capacity to store sequential inputs in designated memory slots and make explicit use of even distantly-preceded information. This approach has particularly been used in the context of long video story comprehension, such as in movies and television series.

Tapaswi *et al.* [142] were the first to adapt and modify a memory network [164] for application in VideoQA. They use the memory to store both video and text features. To facilitate memory read and write operations with enhanced capacity and versatility, Na *et al.* [165] conceptualized a memory network with multiple convolution layers.

Considering the dual-modal information within movie stories, Kim *et al.* [166] incorporated a progressive attention mechanism. The irrelevant video and subtitle segments are first pruned-out using question-guided attention and further filtered using answer-guided attention. The process is repeated multiples times to achieve fine-grained extraction of high-level semantics.

Gao *et al.* [167] proposed a dual-stream framework, so-called Co-Mem, to manage motion and appearance information with a co-memory attention module. The proposed multi-level contextual information allows for the generation of dynamic fact ensembles for a range of questions. Inspired by their work, Fan *et al.* [168] introduced a heterogeneous external memory module (HME) with attentional read and write operations. This was designed to integrate motion and appearance features and learn the spatio-temporal attention simultaneously.

7.3.2.3 Graph-based models

Recently, graph-structured techniques [169] have become increasingly popular for enhancing the reasoning capabilities of VideoQA models. Approaches like HGA [170], B2A [171], and DualVGR [172] construct graphs using coarse-grained video segments while incorporating intra- and inter-modal relationship learning, resulting in solid performance.

To capture object-level details, Huang *et al.* [173] constructed a graph (LGCN) based on objects depicted by their appearance and location attributes, using a Graph Neural Network (GNN) to model the interaction between question-related objects.

Acknowledging the hierarchical nature of video elements in the semantic space, works such as [174]–[176] have integrated hierarchical learning into graph networks. Specifically, Liu *et al.* [177] introduced a graph memory mechanism for relational vision-semantic reasoning from the object level to the frame level. Peng *et al.* [176] progressively linked different-level graphs (object-level, frame-level, clip-level) to understand visual relations (PGAT). Xiao *et al.* [175] developed a hierarchical conditional graph model (HQGA) to merge visual facts from lower to higher-level video elements via graph aggregation and pooling, enabling multi-granular vision-text matching.

7.3.2.4 Transformer-based models

The transformer-based methods can be divided into two categories: *task-agnostic* and *task-specific*. Task-agnostic approaches use unsupervised pre-training on large datasets with universal loss functions then finetuning for tasks like video question answering or video captioning. Details of these strategies are discussed in 7.3.2.4.1. Task-specific techniques directly train models for video question answering. They are reviewed in Section 7.3.2.4.2

7.3.2.4.1 Task-agnostic

Task-agnostic pre-training involves two stages (Figure 7.4). In the first stage, the model is pre-trained on unlabeled large-scale datasets usually involving videos with their associated subtitles. These datasets are easy to acquire automatically from the internet. The main objective of this stage is to learn grounded representations and correlations between visual and textual cues, which form a general-purpose knowledge foundation that can be effectively applied in various downstream tasks. In the second stage, the model is fine-tuned on the downstream task which can be VideoQA, video captioning or multimodal action recognition. More details about vision-language pre-training can be found in [178].

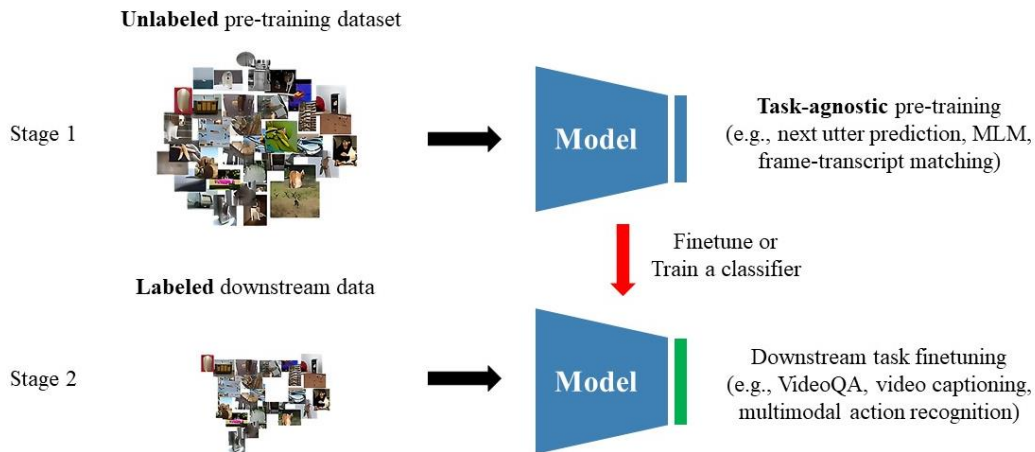


Figure 7.4. Task-agnostic training paradigm.

In [179], the authors propose HERO, a model that leverages a hierarchical transformer architecture to encode multimodal inputs in contrast to the conventional flat BERT-like encoders. The proposed structure consists of a cross-modal Transformer that integrates subtitle sentences with corresponding local video frames and a temporal Transformer that sequentially contextualizes each video frame embedding using the surrounding frames as a global context. The authors introduce four pre-training tasks for HERO: Masked Language Modeling (MLM), Masked Frame Modeling (MFM), Video-Subtitle Matching (VSM), and Frame Order Modeling (FOM). The novelty in this framework primarily lies in VSM and FOM tasks that foster explicit temporal alignment between multimodalities and exploit the sequential nature of video inputs fully. HERO is jointly trained on HowTo100M [137] (narrated instructional videos) and a large-scale TV dataset (comprising varied TV episodes). This combination makes the training dataset more representative of real-life scenarios.

In [180], the authors present CLIPBERT, an efficient framework for end-to-end video-and-language learning, distinguished by two key characteristics. Firstly, in contrast to the dense extraction of video features by most existing methods, CLIPBERT sparsely samples a single or few short clips from full-length videos during each training step. This methodology is based on the hypothesis that these sparse clips encapsulate critical visual and semantic information, as consecutive clips typically exhibit similar semantics from a continuous scene. This approach significantly diminishes memory and computational demands, enabling economic learning from raw video frame pixels and language tokens. The second key feature of CLIPBERT lies in its weight initialization, notably the transfer through pre-training. It adopts 2D architectures like ResNet-50 for video encoding, leveraging the power of image-text pre-training for video-text understanding, along with benefits of low memory cost and runtime efficiency. Through an empirical study, the authors suggest that image-text pre-training indeed contributes positively to video-text tasks, enabling comparable or improved performance on text-to-video retrieval and video question answering tasks. For pre-training the authors use COCO Captions [158] and Visual Genome Captions [181]. Combined, these databases provide a substantial corpus of 5.6 million training instances comprising image-text pairs, derived from a set of 151,000 distinct images.

In [182], authors design a novel dataset for vision-language pre-training called YT-temporal 180M. Unlike previous datasets like HowTo100M, which were limited to instructional videos, YT-temporal 180M encompasses a wide variety of subjects and domains. For pre-training, the authors refine the Masked Language Modeling loss [110] by adding an attention mechanism to filter out unprimed words. They also introduce a novel contrastive frame-transcript matching loss by aligning the video captions with their associative frames. The third and last pre-training loss is temporal reordering of image frames within a video, which allowed the model to learn temporal reasoning. Randomly chosen video frames

are scrambled, and the model is required to reorder these frames based on their correct temporal sequence. The MERLOT-RESERVE approach [183] builds on this framework and adds the audio modality to enhance the performance.

The VIOLET (VIdEO-LanguagE Transformer) model [184] is proposed to improve video modeling for enhanced Video-Language learning. The key-modifications concern the model architecture and pre-training task design. Regarding the model architecture, VIOLET incorporates a Video Swin-Transformer [185], an improvement over the simplistic mean pooling or concatenation methods used on sequences of individual frame features. This transformer explicitly models the video temporality, enabling flexible video-language learning from both videos and static images due to its capacity to handle variable sequence lengths via spatial-temporal self-attention. Concerning the pre-training phase, authors introduce a new, dedicated task, so-called Masked Visual-token Modeling (MVM). The method tokenizes video frames into discrete visual tokens using a pre-trained discrete Variational Auto-Encoder from DALLE [186]. During pre-training, some portions of the video input are masked along both spatial and temporal dimensions, and the model learns to recover these masked patches' discrete visual tokens. The MVM provides an improved approach over Masked Frame/Region Modeling by predicting over a discrete space, avoiding issues of excessive feature dimensions. It also relies on latent visual tokens acquired from a self-reconstruction training procedure, rather than relying on a well-supervised visual backbone. Concerning the datasets, VIOLET is pre-trained using YT-Temporal-180M [182], WebVid-2.5M [187] and ConceptualCaptions3M [125].

7.3.2.4.2 Task-specific models

PSAC (Positional Self-Attention with Co-Attention) [174] is the first framework to apply transformer architecture for VideoQA task. PSAC is comprised of two positional self-attention blocks which substitute the traditional RNN networks for modeling data dependencies. Additionally, a video-question co-attention block (video-to-question and question-to-video attention) is employed. This mechanism enables the model to attend to both critical visual and textual features simultaneously. By doing so, it effectively removes irrelevant video and textual information, ensuring the generation of more accurate answers in the video question-answering tasks. Similarly, [188] apply directly the BERT architecture on multiple-choice VideoQA task by tokenizing the video frames using Faster R-CNN [189].

Engin *et al.* [190] introduce a novel model for scene-based question answering (temporal localization of the answer in TV shows, using video and subtitles). The task requires an understanding of the input dialog and its correlation to the visual cues. The main contribution of this approach is the conversion all input modalities (video, subtitles) into plain text format. The authors employ a visual recognition pipeline to transform raw video data into textual descriptions. The process encompasses four key stages: character recognition, place recognition, object relation detection, and action recognition. The outputs from these stages are aggregated and subsequently compiled into a directed video scene graph. Concerning the subtitles, the authors use BART [191] and SentenceBERT [192] to generate a scene-dialog summary and an episode-dialog summary. Finally, they apply multi-stream text QA techniques to predict the correct answer.

SiaSamRea (self-driven Siamese Sampling and Reasoning) [193] refines the CLIPBERT's [180] sampling strategy and designs a Siamese sampling technique that can generate sparse, yet multiple similar clips, which maintain a consistent global semantic perspective across different starting frames within the same video. Further, they introduce a reasoning strategy called self-driven multimodal learning, designed to fully utilize the Siamese clips. This strategy operates in three steps: firstly, they use sparse and Siamese sampling to obtain both anchor and Siamese clips. These clips are individually cooperated with the text and fed into the model to extract clip-text features. Secondly, an internal contextual interaction is calculated between the anchor clip-text feature and Siamese clip-text features

via a Siamese knowledge generation module. Lastly, they propose a Siamese knowledge reasoning module to reason out the refined soft label. The authors use the pre-trained weights from the CLIPBERT model and finetune it on five different VideoQA datasets.

In [141], authors introduce a novel methodology to automatically generate a large-scale Video Question Answering (VideoQA) dataset, termed as HowToVQA69M. This approach leverages cross-modal supervision, employing transformers that have been trained on an existant text-only question-answering corpus. From this foundation, the system generates video-question-answer triplets using videos and transcribed narrations. The videos are derived from HowTo100M which comprises instructional videos. They also present a new manually annotated open-ended VideoQA benchmark named iVQA. This benchmark is distinctive in that it excludes non-visual questions and provides multiple potential answers for each question, thus extending the richness and complexity of the dataset. The authors apply task-specific pre-training on HowToVQA69M using contrastive learning between a multi-modal video question transformer and an answer transformer. Thanks to contrastive training, they introduce zero-shot VideoQA task to assess the generalization of the model on different datasets. They also apply finetuning on four Factoid benchmarks which are MSRVTT-QA, MSVD-QA [1], ActivityNet-QA [139], and How2QA [179] and prove the effectiveness of task-specific pre-training on HowToVQA69M compared to unsupervised task-agnostic pre-training on unlabeled dataset such as HowTo100M.

7.3.3 Discussion

In this section, we have reviewed different approaches for Video Question Answering task. Monolithic models (section 7.3.2.1), established the fundamental groundwork due to their straightforward design and ease of optimization. These models traditionally leverage recurrence and convolution for processing sequences, but their effectiveness is questionable, as they often yielded sub-optimal results. The limitations of monolithic models, especially in handling long-term dependencies, prompted the incorporation of attention mechanisms. While attention mechanisms contribute towards pruning irrelevant visual and textual cues and enhance the performance to a degree, the reliance on recurrence and convolution in these models remained a setback.

As an alternative, memory-based models (section 7.3.2.2) emerged, relying on an artificial memory to manage and recall past information. However, they still depended on recurrence, which presents the same limitations as with monolithic models. The design and implementation of memory components also adds a level of complexity to these models, creating a new challenge.

We have also reviewed the graph-based models (section 7.3.2.3), which provide a flexible and scalable framework. These models, capable of handling different types of data and network configurations, offered a more intricate representation of relationships in data. However, the challenge lay in constructing graph structures from raw data, which is often computationally intensive. Furthermore, the performance of these models is penalized when the graph representation is not the optimal choice for the data structure.

Given the limitations observed in the aforementioned architectures, our attention turned to transformer-based models for multimodal fusion. Transformers are versatile, accepting any tokenized data type. Their attention mechanisms are particularly adapted for modeling long-range dependencies, enabling the model to attend to all elements in the input sequence.

Task-agnostic pre-training of transformers (section 7.3.2.4.1) has achieved considerable success in the literature and set new performance benchmarks across various tasks. However, these models can be computationally heavy. For instance, the pre-training of the MERLOT model [182] requires approximately 30,000 TPU hours. On the other hand, task-specific models (Section 7.3.2.4.2) are trained directly on the VideoQA task, enabling more efficient capture of task-specific features.

Given these observations, our approach employs task-specific training using transformer-based models. The proposed model leverages the joint-type fusion using a transformer architecture with only two layers, which helps to address the limitations of quadratic complexity often associated with deeper transformers. The reduced depth of our model helps to mitigate computational overheads while still capturing relevant interactions between visual and textual modalities. Additionally, we propose a cross-correlation technique to mitigate heterogeneity between visual and textual modalities, allowing for more effective fusion of information from both modalities. Furthermore, our model is pre-trained on a task-specific dataset, HowToVQA69M [141], which provides it with domain-specific knowledge and improves its performance on video question answering tasks. This pre-training allows our model to better understand the nuances of video data and generate accurate answers to questions posed on video content. We make the hypothesis that our model represents grounding features by jointly learning each modality representation under the constraint of the other.

7.4 Proposed network architecture

The synoptic scheme of the proposed network architecture is presented in Figure 7.5.

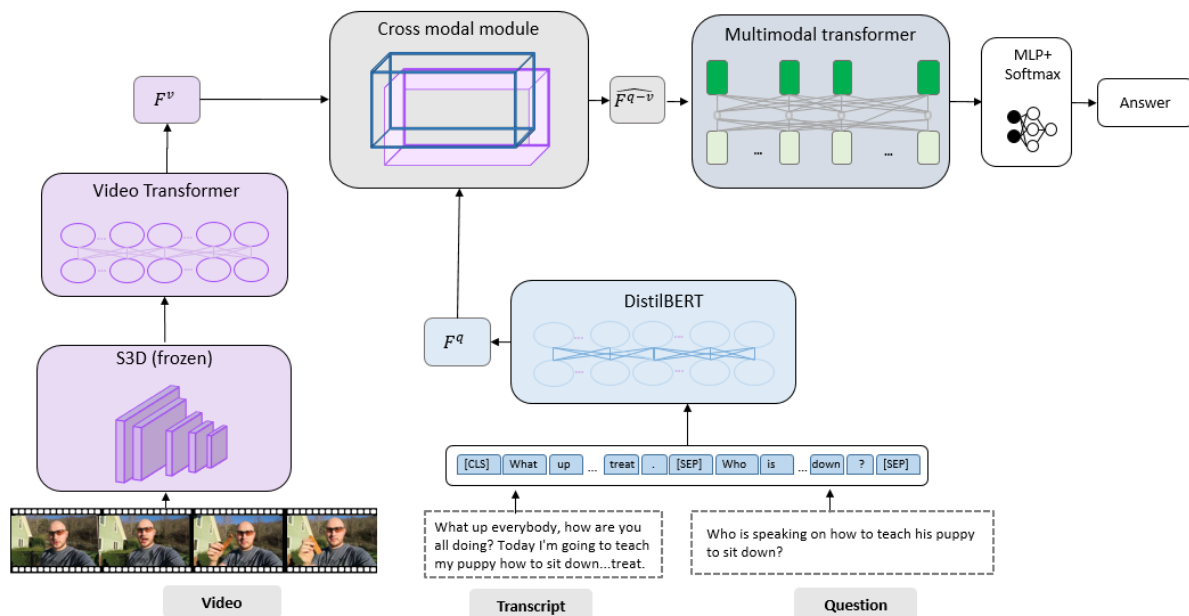


Figure 7.5. The proposed framework for Video Question Answering task.

The various modules involved are the following:

- Feature extraction, which involves pre-processing and tokenization of the input video and text. This step is fundamental as it ensures that the raw video data is transformed into a format suitable for analysis and, likewise, the associated text is structured for effective comprehension by the model.
- Cross-modal module, designed explicitly to minimize the heterogeneity gap between the different modalities inherent to this task. By mitigating this disparity, the model can effectively correlate the video data and the question, forming a more cohesive understanding of the task at hand.

- Multimodal fusion network: This network is based on the transformer architecture, to focus selectively on significant elements from both the video and text, thereby generating a richer and more comprehensive representation of the multimodal data.

Finally, the output of the transformer is then passed to a Multilayer Perceptron (MLP) coupled with a classifier. This final step is dedicated to generating and predicting the most accurate answer to the given question, given the video content.

Let us now detail the various modules considered.

7.4.1 Feature extraction

7.4.1.1 Video processing

To extract the visual representations, the video is uniformly sampled in N fixed length clips of 32 frames. We feed each clip to a S3D model [71]. The idea behind S3D model is to factorize the traditional 3D convolution operation into a 2D spatial convolution operation followed by a 1D temporal convolution operation. The setup can be achieved by conducting two 3D convolutions, where the first (spatial) convolution possesses a filter shape of $[1, k, k]$ and the temporal convolution has a filter of shape $[k, 1, 1]$. This separates the learning of spatial and temporal features in video data, which can help improve the efficiency and performance of the model.

The 3D model has been pre-trained on HowTo100M[137] using the MIL-NCE technique [194]. We retain the feature activations before the final fully connected layer and apply average pooling to obtain a $d_v = 1024$ -dimension vector. The obtained feature vector is denoted as $V = [V_1, \dots, V_N] \in \mathbb{R}^{d_v * N}$ where V_i represents the visual descriptor of the i^{th} video clip. During training, the S3D model weights are frozen to improve efficiency.

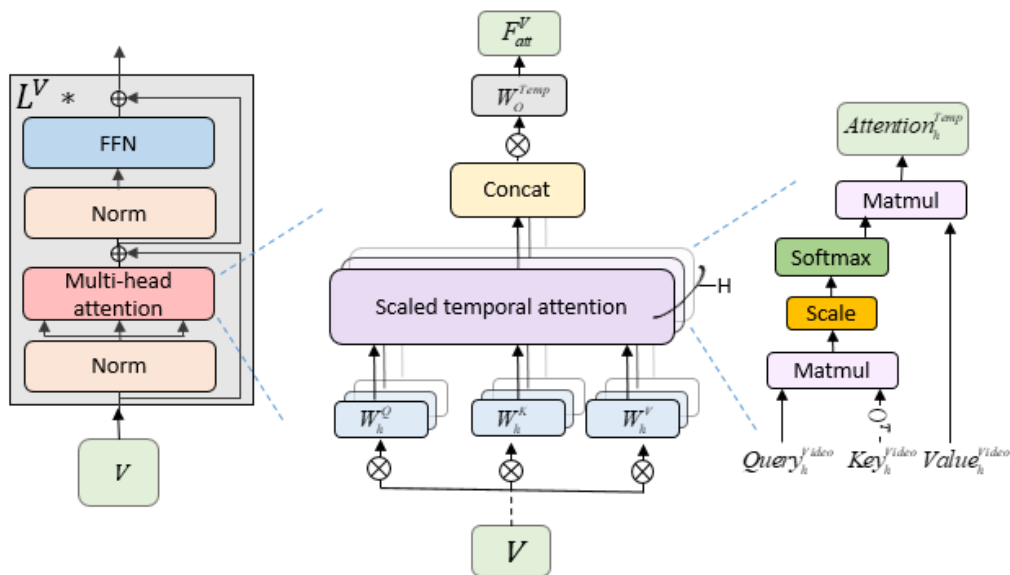


Figure 7.6. The video transformer architecture.

Next, we leverage a video transformer to effectively capture long-range dependencies in video clips. This approach allows us to learn the temporal dynamics of objects, actions and scenes that are inherent in videos. Besides, we can adaptively learn grounded visual features that are specifically optimized for video question answering task without being constrained by pre-extracted features from external models.

To model the dynamic dependencies between clips, we apply temporal attention on the feature vector V as illustrated in Figure 7.6. This approach is motivated by the insight that video data often contains redundant information, and only a limited set of clips contain discriminative information that is relevant for video question answering. To implement this, we apply multi-head temporal attention on the video descriptor V .

For each attention head $h \in \{1, \dots, H^V\}$, we first compute the corresponding $Query_h^{video}, Key_h^{video}, Value_h^{video}$ and as follows:

$$Query_h^{video} = VW_h^Q, Key_h^{video} = VW_h^K, Value_h^{video} = VW_h^V \quad (7.5)$$

where, W_h^Q, W_h^K and W_h^V are three learnable matrices.

The temporal attention for a given attention head is computed as:

$$Attention_h^{Temp} = softmax\left(\frac{Query_h^{video}(Key_h^{video})^T}{\sqrt{d_{Temp}}}\right)Value_h^{video} \quad (7.6)$$

The result of all attention heads are then concatenated and once again projected using a learnable matrix W_O^{Temp} as follows:

$$F_{att}^V = Concat(Attention_1^{Temp}, \dots, Attention_{H^V}^{Temp})W_O^{Temp} \quad (7.7)$$

Finally, a feed forward network (FFN) with linear projection followed by GeLU activation function [195] and layer normalization is used to project the feature vector. The spatio-temporal features are denoted by $F^v = [F_1^v, \dots, F_N^v] \in \mathbb{R}^{d \times N}$, where d is the dimension of the projection space.

7.4.1.2 Text processing

When audio is available in the dataset, we utilize an ASR (*Automatic Speech Recognition*) model to extract the associated transcript. In our implementation, we have retained the Whisper library [196], an end-to-end transformer-based model that exhibits human-level robustness in English speech recognition, even in the presence of background noise and reverberation. We pre-process the dataset offline to speed up the training process (Figure 7.7).

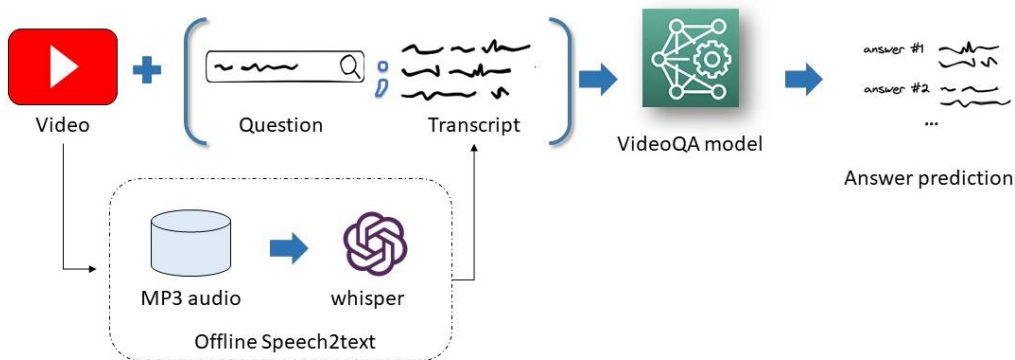


Figure 7.7. Overview of text processing framework.

The text input is first tokenized using the WordPieces tokenizer [197], a sub-word segmentation algorithm with a 30,000 token vocabulary. When the audio is available, we append the transcript to the

question as following: “[CLS] transcript. [SEP] question. [SEP]”, as suggested in [116] (Figure 7.7). The [CLS] token marks the beginning of the sentence and its final hidden state is used as a summary representation of the whole sequence. The [SEP] token serves as a boundary marker, allowing the model to distinguish between audio transcripts and textual questions and encode them independently. Each token is then fed to DistilBERT [198]. DistilBERT is an efficient, lightweight version of BERT, which is trained under low latency constraints. We use the activations of the last layer of DistilBERT to obtain a 768-dimensional feature vector which is then passed to a feed forward network, similarly to the video projection. The text embedding is denoted as $F^q = [F_1^q, \dots, F_T^q] \in \mathbb{R}^{d \times T}$, where T is the number of tokens in the text.

7.4.2 Cross-modal module

Modeling video-text dynamics within and across modalities is an extremely challenging task. To mitigate this problem, we have developed a cross-modal correlation module that efficiently accounts both intra and inter-modal relationships between modalities. Figure 7.8 illustrates the architecture of the proposed module.

We consider the cross-correlation matrix τ (Eq. (7.8)) that aims at modeling the relationships between the various visual and textual modalities involved:

$$\tau = F^{qT} W F^v \quad (7.8)$$

where $W \in \mathbb{R}^{d \times d}$ is a learnable matrix.

A high coefficient of the cross-correlation matrix τ means that the corresponding video and text features are highly relevant. We generate the cross-correlation video-text (resp. text-video) weights by column-wise softmax over τ and τ^T , respectively. This technique allows learning more discriminative representations for each individual modality, constrained by the other one. Formally, we compute the video-conditioned text features as:

$$F^{q-v} = F^q \text{softmax}(\tau^T) \quad (7.9)$$

Similarly, the text-conditioned video features are defined as:

$$F^{v-q} = F^v \text{softmax}(\tau) \quad (7.10)$$

To prevent information loss in the cross-correlation stage, we have adopted the dense skip connection technique. The reweighted features F^{q-v} and F^{v-q} are thus added to the original modality-specific representation.

$$\widehat{F^{q-v}} = \tanh(F^{q-v} + F^q) \quad (7.11)$$

$$\widehat{F^{v-q}} = \tanh(F^{v-q} + F^v) \quad (7.12)$$

The obtained features are further exploited in the multi-modal fusion module, as described in the following section.

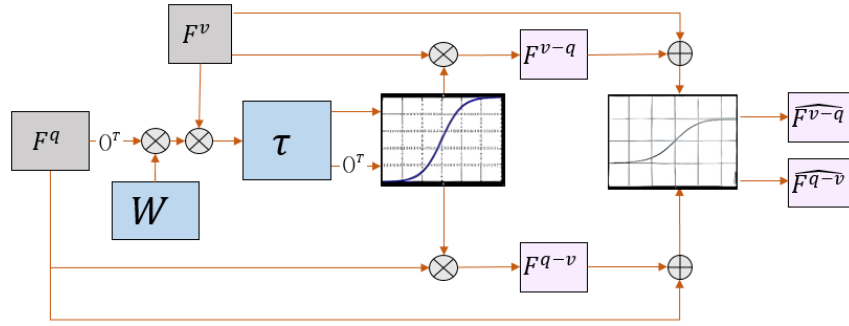


Figure 7.8. Cross-modal correlation module.

7.4.3 Transformer-based multimodal fusion

In contrast to recurrent neural networks, transformers are order-insensitive due to their self-attention mechanism. This means that the order of tokens in the input does not impact the model's ability to process them. To address this limitation, we have considered the positional encoding introduced in [4], as described in the following equations:

$$pos_{2k} = \sin\left(\frac{i}{10000^{2k/d}}\right) \quad (7.13)$$

$$pos_{2k+1} = \cos\left(\frac{i}{10000^{2k/d}}\right) \quad (7.14)$$

where i represents the position of the token in the input sequence, k represents the index of the positional encoding, and d represents the dimensionality of the video-text embeddings. This encoding explicitly retains information about the token positions in the input sequence, allowing thus the transformer model to capture and take into account the positional information.

Additionally, we incorporate a learned modality embedding layer to differentiate between the two modalities, video and text. The modality embedding is added to each token as an additional feature, and it is learned during the training process. The modality embedding is calculated as follows:

$$mod_m = \text{softmax}(\text{linear}(\text{onehot}(m))) \quad (7.15)$$

where m represents the modality of the token (either video or text), $\text{onehot}(m)$ represents the one-hot encoding [199] of the modality (variables are represented as binary vectors where each vector has a 1 in the position of the corresponding category and 0 in all other positions). This modality embedding allows the model to explicitly differentiate between the two modalities and capture their distinct characteristics.

The representations of the video and the text are computed as follows.

$$\widetilde{F}_k^{q-v} = dp(\widetilde{F}^{q-v} + pos_k + mod_q) \quad (7.16)$$

$$\widetilde{F}_k^{v-q} = dp(\widetilde{F}^{v-q} + pos_k + mod_v) \quad (7.17)$$

where $mod_q \in \mathbb{R}^d$, $mod_v \in \mathbb{R}^d$ represent the learnt modality embeddings; and $[pos_1, \dots, pos_{t+c}] \in \mathbb{R}^{d \times T+N}$ are positional encodings. dp is the dropout layer.

The input to the transformer $F^{qv} \in \mathbb{R}^{d \times T+N}$ is the concatenation of \widetilde{F}_k^{q-v} and \widetilde{F}_k^{v-q} .

The transformer layers consist of an attention sublayer followed by a position-wise feed-forward layer. The attention sublayers employ H attention heads. To obtain the sublayer output $O \in$

$\mathbb{R}^{seq_length*d}$ ($seq_length = T + N$), we concatenate the results from each head and apply a linear projection. Each attention head operates on an input sequence $X \in \mathbb{R}^{H*seq_length*d_head}$ and computes the attended feature $Z \in \mathbb{R}^{H*seq_length*d_head}$ as follows.

$$z_i = \sum_{j=1}^{seq_length} \alpha_{ij}(x_j W^V) \quad (7.18)$$

The weight coefficient α_{ij} is calculated using a softmax function.

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{t+c} e_{ik}} \quad (7.19)$$

where ,

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_{head}}} \quad (7.20)$$

with $W^Q, W^K, W^V \in \mathbb{R}^{d*d}$ learnable matrices and d_{head} denoting a scaling factor.

Finally, the fused video-text representation is obtained as:

$$F = W_{vq} dp(Q_1) + b_{vq} \quad (7.21)$$

where W_{vq}, b_{vq} are learnable parameters and Q_1 is the multimodal contextualized embedding of the [CLS] token in the input text as in [110]. We use the softmax function to predict the correct answer from the vocabulary of predefined answers.

7.5 Rephrasing attacks

7.5.1 Problem formulation

The objective of adversarial attacks is to fool the learned model by manipulating the input provided to it. This is not only important to test the vulnerability of DL models to security threats but also to evaluate its robustness in real-world scenarios. Adversarial attacks have been first introduced in the image domain for object recognition [200]–[202], then attracted many follow-up efforts in other domains including natural language processing (NLP). Text attacks are more challenging due to different reasons: (1) Small changes in the image are unperceivable by humans while text changes can be easily identified; (2) The semantics of the image is not changed by small perturbations. In contrast, even minor text manipulations can affect the general meaning of a sentence.

A successful attack should take such considerations into account, in order to be able to fool the DL model without changing the human judgement. Adversarial attacks can be categorized into two classes. A first one concerns the so-called *white box attacks*: in this setting, the attacker has access to the model information including input-output data, model architecture, parameters, loss functions and activation functions. The adversarial data is adjusted to maximize its influence on the classifier while keeping an imperceptible change. Most approaches use the gradient information of the loss with respect to the input to build the attack. In [203], authors use fast gradient sign method (FGSM) [201] by identifying the words with the most significant contribution to classification task. Specifically, they compute the cost gradient of training examples using backpropagation and assign the contribution of each item with respect to the magnitude of the cost gradient. Jacobian Saliency Map Adversary (JSMA)-based methods [204], [205] build adversarial perturbations using forward derivatives.

In the case of the second family of methods, called *black box attacks*, the attacker has only access to input-output data. This approach uses heuristic methods or iterative queries to perform the attack.

In [206], authors distract the textual input by appending meaningless sentences at the end of the paragraph. Such perturbations are crafted by iteratively querying the model until the output changes. In [207], various strategies are applied to affect the model’s performance such as random swap (transposing neighbor words), random deletion, stop-word dropout, paraphrasing as well as grammar and keyboard errors. In [208], [209], the important tokens are identified based on a scoring system which measures the degree of perturbation of the model’s output. The selected tokens are then modified using four techniques: delete, replace, swap and add. In [210], [211], authors generate semantically equivalent adversaries (SEA) to fool the model. Such approaches generate paraphrases and compare the model’s prediction with the original sentences. Other works [212], [213] leverage generative adversarial networks (GANs) [214] to generate adversarial examples by searching for the neighbors of the input data in the latent space.

The output of the adversarial attacks can be *targeted*, meaning that the attacker maps the output to a desired value, or *untargeted* in which case the attacker cares only about producing incorrect output. For multimodal attacks, there has been some work on image captioning [215], optical character recognition [216] and image question answering [217]. To the very best of our knowledge, this is the first work to consider adversarial attacks issues under the framework of video question answering methods. The framework of our approach is illustrated in Figure 7.9.

7.5.2 Methodology

Our objective is to verify the importance of the building elements of our pipeline and test their respective contribution to the model prediction. To allow a fair comparison, the same model-independent attacks are applied on the different models. For this reason, we apply untargeted black-box attacks, meaning that we do not enforce any specific results. We use an automatic method to generate the rephrased questions without additional human intervention, which is more scalable in real-world environments. To this purpose, we have retained the BART approach [191], which is a sequence-to-sequence NLP model that uses a BERT-like encoder (i.e., bidirectional encoder) and a GPT-like decoder (i.e., left-to-right decoder). BART is pre-trained in an unsupervised manner using general objectives such as text corruption with random noise and text shuffling. The model is originally applied to sequence generation and machine translation tasks. The model is fine-tuned for text rephrasing purposes. The pre-trained model is directly used as a sequence-to-sequence model. At each time step, the model computes the probability of each word in the vocabulary to be the likely next word. Then, the next word is picked based on three decoding methods: (1) random sampling: we randomly choose the next word w_t according to its conditional probability distribution. (2) top-K sampling [218]: we only sample the K high probability words from the distribution. (3) top-p (nucleus) sampling: we sample from a set of words whose cumulative probability exceeds p .

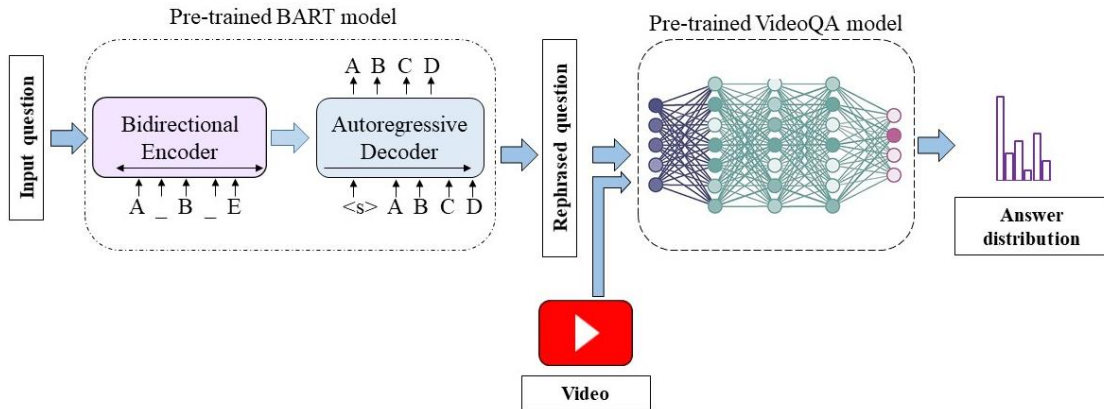


Figure 7.9. Rephrasing attacks on Video Question Answering model.

For training, we use three datasets: Quora [219] (400k training samples), MSRP [220] (13M training samples) and PAWS [221] (108k training samples). The original data is filtered to ensure more diversity as follows. First, the sentence pairs that present more than 80% unigram overlap are removed. This first step minimizes the chance to copy the original sentence. We use Siamese BERT [192] to remove the question pairs with low semantic similarity. For MSRP and Quora, we only select the sentences that are rephrases to each other. Finally, the trained model is applied on the test set of MSVD-QA.

Some examples of rephrased questions are provided in Table 1. In order to compare the differences between the two datasets (original and rephrased) we compute the GLEU score [197] which is more suitable for single sentences. GLEU is a variant of the BLEU score that assigns more weight to n -grams that are changed from the source. Specifically, the GLEU score is the minimum of recall (ratio of the number of matching n -grams to the total number of n -grams in the original question) and precision (the ratio of the number of matching n -grams to the total n -grams in the rephrased question). The GLEU score range is between 0 (no matches) and 1 (all match). We have obtained a GLEU score of 0.5638.

Table 7.2. Examples of rephrased questions from MSVD-QA dataset.

Original question	Rephrased question
Who is on an ambulance stretcher	Who is riding an ambulance stretcher?
What are school aged children doing?	What is a group of teenagers doing?
How many elephants are spraying water on themselves?	How many elephants are spraying water on themselves with their trunks ?
What is the best way to cut potato into pieces with a knife?	Who is cutting into pieces a potato with a knife?
What does a man pick up a card from?	What does a man pick a card up from?
What is climbing?	What is climbing?

7.6 Experimental evaluation

The experimental evaluation has been carried out on the publicly available datasets HowToVQA, MSVD-QA and MSRVT-T-QA [1].

7.6.1 Datasets

HowToVQA: Under the framework of a *pre-training, then fine-tuning* paradigm, we have trained the model on the HowToVQA 69M task-specific dataset. HowToVQA 69M is today the largest VideoQA available dataset, with over 69 million video question-answer triplets. The videos have been extracted from HowTo100M, which was originally designed for video captioning purposes. The question-answer pairs are automatically generated from the transcribed speech using two transformers. We randomly select 164148 training samples to reduce memory and computational requirements.

MSVD-QA: For fine-tuning, we have retained the popular MSVD-QA VideoQA dataset, which represents a smaller dataset automatically derived from MSVD. The dataset contains 1970 video clips, each 10-seconds long and featuring a single activity, along with 50,505 open-ended question-answer pairs that require an understanding of the video content to answer correctly. The questions cover a wide range of topics and were created automatically from the existing captions in MSVD dataset. The answer vocabulary contains 1852 training answers.

MSRVTT-QA: We also used the MSRVT-T-QA dataset. This dataset includes 10,000 video clips, each 20-seconds long, and 200,000 open-ended question-answer pairs covering a diverse range of topics. The questions are also automatically generated from the available captions. Unlike the MSVD-QA dataset, the MSR-VTT-QA dataset also includes the audio channel. We do not apply pre-training on MSRVT-T-QA as the dataset is small enough to perform the training from scratch.

7.6.2 Implementation details

For pre-processing, we uniformly sample the video into $N = 20$ clips. Similarly, we set the maximum number of tokens in the input question to 20. We project the video features and text features into a

common embedding space of size $d=512$. We process the contextual features of the video using $L_v = 2$ layer transformer with $H^v = 8$ attention heads and a scaling factor $d_v = \frac{d}{H^v} = 64$. Regarding the transcript, we utilize the Whisper ASR model to extract the speech from the video. We apply the Whisper model on the entire video rather than on individual clips, as people commonly mention key objects or actions before or after they are shown in the video. We set the maximum number of tokens in the transcript to 20.

For the multimodal transformer, a number of $H = 8$ attention heads has been retained. In this setting, the scaling factor d_{head} is the fraction of the embedding size over the number of heads $d_{head} = \frac{d}{H} = 64$. To train the rephrasing model BART, we select the high probability words based on top-K and p-sampling strategies. We set $K=50$ and $p=0.95$.

The loss function of the proposed model is the sum of the cross entropy loss and the masked language modeling (MLM) loss. The MLM objective is to predict a randomly masked word from a predefined vocabulary of 30K words. MLM loss is the negative log-likelihood for masked words. Specifically, we randomly select with a probability of 15% all WordPiece tokens in each question. Once the token is selected, the data generator replaces the token with a special token [MASK] 80% of the time, a random token 10% of the time, and the same token 10% of the time. The goal of this procedure is to influence the model to maintain a contextual representation of each input token, since it does not know which words will be predicted.

A cosine annealing learning rate schedule has been used, with initial values of 10^{-4} for training on both HowToVQA and MSRVTT-QA, and 10^{-5} for fine-tuning on MSVD-QA respectively. For optimization, we have adopted the Adam approach with batch size of 16 for pre-training and 32 for fine-tuning. The training process has been run on 2 NVIDIA GeForce RTX 2080 GPUs and for 20 epochs.

The final model is selected according to the best performance on the validation set.

7.6.3 Ablation studies

7.6.3.1 Ablation studies on MSVD-QA

To investigate the effectiveness of each component of the pipeline, we have compared the performance of different baselines on both original and rephrased datasets using the MSVD-QA dataset. More precisely, we have first retained the following four baseline methods:

- *Early fusion* which concatenates the video features from S3D and text representations and feeds them directly into a fully connected layer to predict the correct answer.
- *Cross modal module only (CMM)* that learns inter-modal representations of each modality under the constraint of the other (*cf.* Section 3.2).
- *Multi-modal transformer (MMT)* which feeds the representations of the pre-trained model to a joint transformer and neglects the cross modal module (*cf.* section 3.3).
- *CMM+MMT* which uses the cross modal module in conjunction with the multimodal transformer trained from scratch on MSVD-QA.
- *CMM+MMT+PF*, which is the same as the latter but pre-trained on a subset of HowToVQA 69M then fine-tuned on MSVD-QA (PF).

Let us underline that none of the above-mentioned baseline methods uses the video transformer. Instead, they extract visual features from the frozen layers of the S3D model.

In a second part, we have considered the complete framework, integrating the transformer-based approaches, with two variants:

- *CMM+MMT+VT*, which is our framework including the video transformer (VT) trained from scratch on MSVD-QA,
- *CMM+MMT+VT+PF*, which is our framework pre-trained on a subset of HowToVQA69M and fine-tuned on MSVD-QA.

In order to evaluate the performances, we have adopted the accuracy metric as the answers do not exceed several words. The accuracy represents the ratio of the correct predictions with respect to the total number of input samples. The obtained results are summarized in Table 7.3.

The following conclusions can be drawn: (1) The lowest score is obtained by directly concatenating video and text representations. This behavior can be explained by the heterogeneous nature of the two modalities involved which are pre-trained with different tasks/datasets. (2) Cross-correlation technique yields more grounded representations as features are learnt under the constraint of the other modality, with a 3.72% improvement in accuracy. (3) The video transformer improves the results by 0.58% when the model is trained from scratch and by 1.38% when the model is pre-trained on HowToVQA. This result proves the effectiveness of the video transformer architecture in capturing long-range dependencies in videos. Consequently, the model's ability to process complex video data and produce more accurate results is improved. (4) The best results are obtained using the full pipeline, which integrates extensive inter-modal interactions. (5) Pre-training on large task-specific datasets effectively optimizes the weights of the proposed architecture. (6) Our approach is more robust to rephrasing attacks than the transformer-only architecture. This is due to learning-conditioned features as opposed to simple concatenation.

Table 7.3. Ablation studies on MSVD-QA. Acc1 represents the performance on the original dataset. Acc2 represents the performance on the rephrased dataset.

Methods	ACC1	ACC2
Early fusion	27.33%	21.31%
CMM	31.05%	25.81%
MMT	37.88%	33.47%
CMM+MMT	38.96%	33.87%
CMM+MMT+PF	43.58%	39.42%
CMM+MMT+VT	39.54%	34.51%
CMM+MMT+VT+PF	44.96%	41.09%

Figure 7.10 provides examples of results on the MSVD-QA dataset, demonstrating both original and rephrased questions. In general, the model's predictions on the test set and rephrased set align well with the ground truth. However, as shown in example (f), the adversarial attack may sometimes alter the results. Interestingly, rephrasing the question can also improve the accuracy of the model, as shown in example (h), where rephrasing has eliminated additional noise in the question by focusing only on the relevant textual cues. These findings suggest that data augmentation using rephrased samples could potentially yield better results by adding more training samples and reducing inherent biases during training.

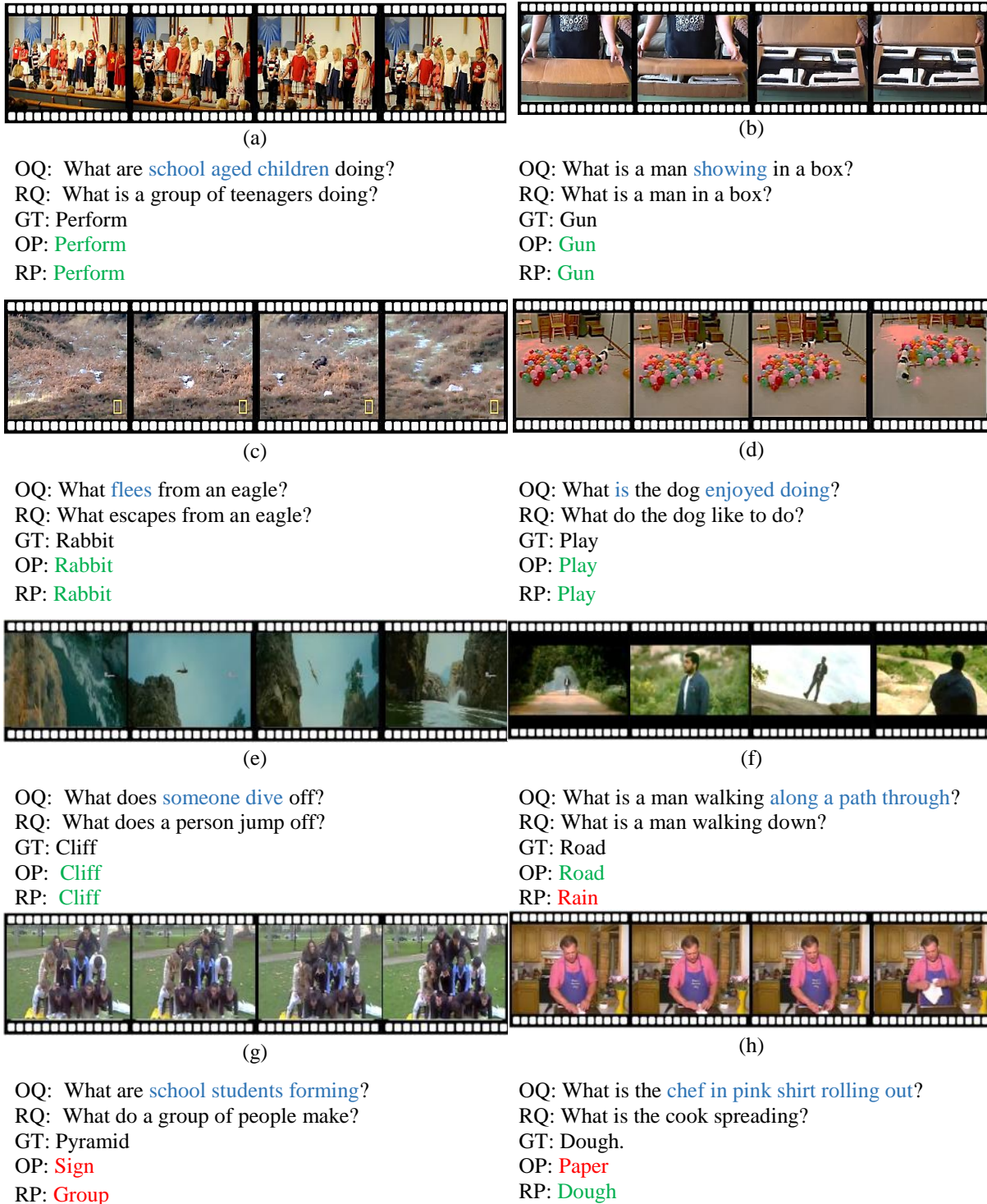


Figure 7.10. Examples of results of our approach on the MSVD-QA dataset, with both original and rephrased questions. OQ: Original question; RQ: Rephrased Question; GT: Ground Truth; OP: Prediction of the model to the Original question; and RP: Prediction of the model to the rephrased question.

7.6.3.2 Effect of the transcript input

We have also investigated the impact of incorporating transcript information into the question using the MSRVTQ-QA dataset, as described in section 7.4.1.2. We have compared the performance of our model trained with only the question input versus the model trained with both the question and transcript as input text. To this purpose, we have manually annotated two subsets of the test set. The first subset is referred to as "vision-text" samples, which require both video and transcript information to answer questions, such as identifying the main topic of a discussion. The second subset is called "vision-only" samples, which can be answered based on visual cues only, such as identifying an action or object in the video. We have identified videos where the transcript conveys semantic meaning and labeled them as "vision-text" samples (Figure 7.11.(c)). In contrast, videos lacking a transcript (Figure 7.11.(e)) or with an incomprehensible transcript (Figure 7.11.(h)) were classified as "vision-only" samples. Our manual annotations resulted in 663 vision-only videos and 552 vision-text videos being labeled. The results obtained are reported in Table 7.4, where Acc_{test} is the accuracy on the whole test set, Acc_{v-t} represents the accuracy on the "vision-text" subset and Acc_v represents the performance on the "vision-only" subset. The results show that incorporating the transcript information improves accuracy by 1.86% on the test set. Furthermore, the improvement is even greater for vision-text samples (2.1% gain) compared to vision-only samples at 1.65%. However, the results are globally better for vision-only samples. We believe that including the transcript as an additional modality input could potentially yield even better results, at the cost of an increased complexity.

Table 7.4. Comparison of the effect if the transcript input on MSRVTQ-QA dataset.

Methods	Acc_{test}	Acc_{v-t}	Acc_v
Model trained with only the question as text input (MQ)	40.02%	38.90%	44.23%
Model trained with both question and transcript as text input (MQT)	41.88%	41.00%	45.88%

In Figure 7.11, we present some examples of results from the MSRVTQ-QA, showcasing the performance of two models: (MQ) trained with only the question as input text, and (MQT) trained with both the question and transcript as input text. In general, the generated answers are accurate and well-aligned with the video content. Both MQ and MQT are able to correctly identify the most salient visual features in the video, as seen in examples (a) and (b). However, when the ASR transcript is available, MQT outperforms MQ in identifying the main topic or subject being discussed, such as in examples (c) and (d) where MQ fails to generate the correct answers due to the lack of relevant visual cues. Nonetheless, there are some discrepancies in the generated answers. For instance, in example (g), MQT predicts "perfume" which is correlated to the word "smell" in the question and the word "perfume" in the generated transcript, which is not the correct answer. In example (h), both models rely solely on the visual element "helicopter", disregarding the actual question. We attribute these discrepancies to the quality of the automatically generated transcript and the clarity of the question.

Overall, the results suggest that incorporating the transcript in the input text can enhance the accuracy of generated answers in video-based question answering tasks, compared to using only the question and video as input.



(a)

ASR: What up everybody, how are you all doing? Today I'm going to teach my puppy how to sit down. Make sure you're gonna grab a good treat.

Question: Who is speaking about how to teach his puppy to sit down ?

GT: Man

MQ: Man

MQT: Man



(b)

ASR: She's part of the family. Would you like to hold her? Of course, please. Come here. Look, Bella. You smell so clean. Check this out guys.

Question: What does a man hold?

GT: Dog

MQ: Dog

MQT: Dog



(c)

ASR: Well, this is the mother of all earthquake faults. It can pack wallop 30 times that of the San Andreas fault. So forget all the Hollywood hype about the San Andreas fault. We're talking about an earthquake a night.

Question: What is a man talking on tv about?

GT: Earthquake

MQ: Color

MQT: Earthquake



(d)

ASR: In the film, McConaughey said that he was in a Tesseract. Let me ask you something, how do I phrase this? What the F is a Tesseract?

Question: What do two men discuss?

GT: Movie

MQ: Video

MOT: Movie



(e)

ASR: ∅

Question: What is woman doing?

GT: Play

MQ: Play

MQT: Play



(f)

ASR: ∅

Question: What is being displayed?

GT: Ship

MQ: Ship

MOT: Ship



(g)

ASR: This smells awesome. Oh no, it doesn't. This does smell like a middle schooler's perfume though, right? It does, yeah. It's like a jello shot.

Question: What are people smelling?

GT: Drink

MQ: Women

MQT: Perfume



(h)

ASR: The new fronts and the fighting bring fresh hill and the groaned.

Question: What is a helicopter being shot at by ground forces in?

GT: Syria

MQ: Helicopter

MOT: Helicopter

Figure 7.11. Examples of predictions on MSRVTT-QA dataset.

7.6.4 Comparison with state-of-the-art

We have compared our approach to various state of the art methods on the MSVD-QA and MSRVTT-QA datasets [1]. Table 7.5 summarizes the accuracy of the different VideoQA models retained for comparison. Our evaluation includes monolithic models including ST-TP [140], AMU [1], and HCRN [222], graph model B2A [171], memory-based models including Co-Mem [167], and HME [168] and transformer-based models including CLIPBERT [180], and CoMVT [223].

The proposed method achieves the highest accuracy scores on both MSVD-QA (44.96%) and MSRVTT-QA (41.88%) datasets. In particular, it outperforms the state of the art CoMVT model by 2.36% on MSVD-QA and 2.38% on MSRVTT-QA. Interestingly, CoMVT is pre-trained on a larger, task-independent dataset (HowTo100M). It uses four transformer blocks to model intra- and inter-model dynamics, while we use a simple weight matrix followed by a 2-layer transformer. This demonstrates the importance of task-oriented pre-training and the effectiveness of our model, which minimizes the required computational effort.

Table 7.5. Comparison with state-of-the-art models on MSVD-QA and MSRVTT-QA

Methods	MSVD-QA	MSRVTT-QA
ST-TP [140]	31.3%	30.9%
AMU [1]	32.0%	32.5%
Co-Mem [167]	31.7%	31.9%
HME [168]	33.7%	33.0%
HCRN [222]	36.1%	35.6%
B2A [171]	37.2%	36.9%
CLIPBERT [180]	-	37.4%
CoMVT [223]	42.6%	39.5%
Our model	44.96%	41.88%

7.7 Conclusion

In this chapter, we have proposed a novel multimodal framework for video question answering. The proposed system is based on reciprocally constrained, cross-correlation conditioning of visual and textual features. Our system also integrates attention mechanisms using a multimodal, transformer-based approach to capture complex inter-modal dynamics. Additionally, we have used a video transformer incorporating temporal attention to learn contextual features of the video rather than relying on pre-extracted frozen features from external models. Ablation studies demonstrate the importance of each composing block of the approach. We have also proved the effectiveness of our pipeline by testing the robustness of the model to rephrasing attacks. Furthermore, we have investigated the importance of the transcript modality in providing the correct answer while maintaining the same model. The experimental results obtained show that the proposed framework achieves high accuracy scores, with 44.96% and 41.88% accuracy on the MSVD-QA and MSRVTT-QA datasets, respectively. In addition, it outperforms previous state-of-the-art methods (by 2.36% and 2.38%, respectively).

For future work, we plan to extend our video question answering framework to incorporate audio features and explore their impact on model performance. Additionally, we intend to apply our model on a real video-question platform to conduct a subjective system evaluation with user feedback, in order to further validate the effectiveness of our approach.

8 VIDEO CAPTIONING

Abstract: In this chapter, we introduce a novel end-to-end multimodal video captioning framework based on visual and textual fusion. The proposed approach integrates a modality-attention module, which captures the visual-textual inter-modal relationships using cross correlation. Further, we integrate temporal attention into the features obtained from a 3D CNN to learn the contextual information in the video using task-oriented training. In addition, we incorporate an auxiliary task that employs contrastive loss to enhance the model's generalization capability and foster a deeper understanding of inter-modal relationships and underlying semantics. The task involves comparing the multimodal representation of the video-transcript with the caption representation, facilitating improved performance and knowledge transfer within the model. Finally, we use a transformer architecture to effectively capture and encode the interdependencies between the text and video information using attention mechanisms. During the decoding phase, the transformer allows the model to attend relevant elements in the encoded features, effectively capturing long-range dependencies and ultimately generating semantically meaningful captions. The experimental evaluation, carried out on the MSRVTT benchmark dataset, validates the proposed methodology which achieves BLEU4, ROUGE and METEOR scores of 0.4408, 0.6291 and 0.3082, respectively. When compared to the state of the art methods, our system shows superior performances, with gains in performance ranging from 1.21% to 1.52% across the three metrics considered.

Keywords: Video captioning, multimodal learning, cross correlation.

8.1 Introduction

Video captioning refers to the task of generating descriptive, natural language text from the input video, encapsulating the semantic and taking into account the dynamic elements that are present in the video sequence (Figure 8.1). This task extends beyond simple image captioning as it involves a comprehension of temporal sequences and relationships, in addition to static objects and their interactions. Thus, video captioning requires the full understanding of a video's context, objects, scenes, and actions, further mapping these aspects into a coherent natural language sentence or paragraph.

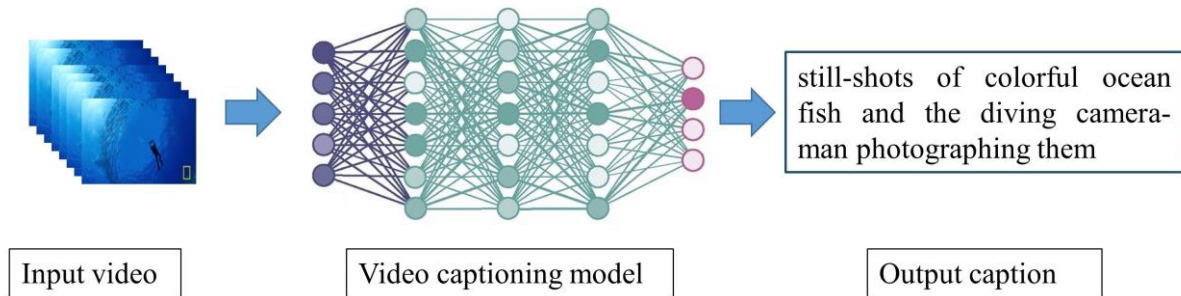


Figure 8.1. Video captioning problem.

This chapter is organized as follows.

In Section 8.2, we briefly recall the challenges of automatic video captioning within the framework of TV archive indexation.

In Section 8.3, a comprehensive state-of-the-art review is proposed. The state of the art methods are classified into template-based and deep-learning based approaches, the latter being further subdivided into visual-based and multimodal methodologies. We also lay out the commonly-used benchmark evaluation metrics such as BLEU, ROUGE, and METEOR.

Section 8.4 introduces the proposed architecture, which is built upon an encoder-decoder paradigm. The process starts with the extraction of visual and textual features, followed by the application of a modality-attention module similar to that used in the video question answering framework. To encode the multimodal features, a transformer encoder is utilized. The subsequent generation of the decoded caption is then achieved through a transformer decoder, which makes use of the previously generated tokens and encoder information to guide the training process.

In Section 8.5, we outline the model's training objectives. A masked language modeling technique is employed to facilitate the learning of robust textual representation. We incorporate an auxiliary task based on a contrastive loss between the multimodal video-transcript and caption features, thereby aligning the input more closely with its corresponding caption. Cross-entropy is applied on the decoded tokens to evaluate the caption generation.

Section 8.6 details the experimental evaluation performed and the results obtained. In order to assess the importance of the building blocks within our framework, extensive ablation studies are carried out, and our approach is compared with previous state-of-the-art methods.

Finally, we conclude the chapter in Section 8.7, outlining potential future directions for research in this field.

8.2 Application to TV archive indexing

The process of summarizing video content constitutes a crucial component of the responsibilities borne by documentalists within the television broadcasting industry. This typically involves the creation of natural language descriptions for each video shot (Figure 8.2). However, this task is not without its inherent difficulties, with the process being labor-intensive, time-consuming, and subject to potential human bias. In this chapter, we propose an automatic technique that seeks to alleviate these challenges by synthesizing video content using advanced video captioning methodologies. Our method automates the creation of comprehensive, contextual video summaries, thereby offering a potential resolution to the time-intensive nature of manual video summarization. Furthermore, the inherent objectivity of an automated system addresses the issue of human bias, providing consistent and impartial summaries. Crucially, the application of video captioning for archive indexing could significantly enhance content retrieval processes, enabling faster and more efficient location of relevant archived footage. In essence, this video captioning-based approach presents a substantial stride forward in streamlining the documentalist's workflow and augmenting the operational efficiency of TV broadcast archives.

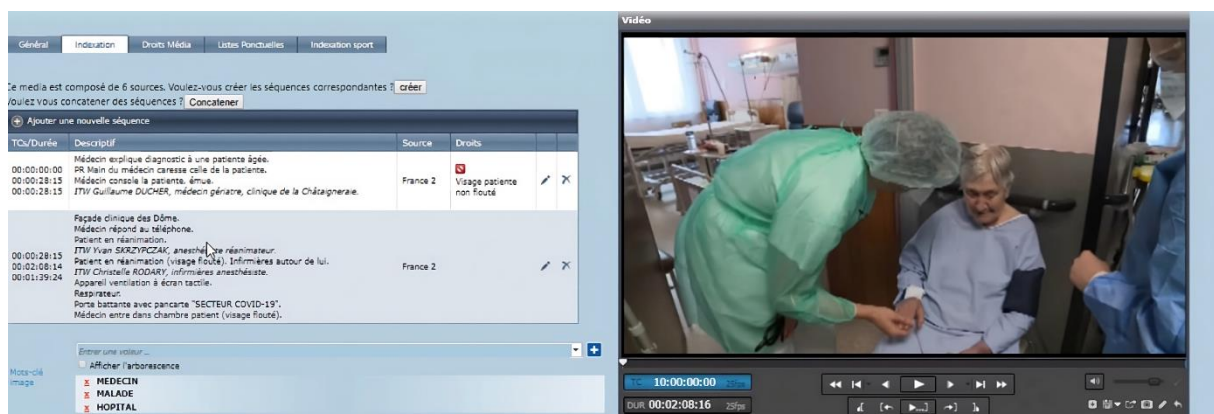


Figure 8.2. An archive indexing page. The field « descriptif » represents the natural language description of the video content at the shot level. Source: DALET.

8.3 Related work

Video captioning is a well-explored domain within the literature with initial methods dating back to 2002 [224]. This section reviews these methodologies, beginning with early template-based techniques (Section 8.3.1). We then transition into an analysis of deep learning approaches, classifying them into visual-based methods (Section 8.3.2.1) and multimodal strategies (Section 8.3.2.2). Finally, we present the various evaluation metrics used in the literature. We do not discuss the video captioning datasets as most of them were already outlined for video question answering (section 7.3.1).

8.3.1 Template-based approaches

Building on the achievements in image recognition and activity recognition, one straightforward method involves the conversion of detected outputs into a coherent sentence using a template, which ensures grammatical accuracy. This template-based language process initially dissects sentences into fragments, such as the subject, verb, and object, guided by specific grammar rules. Each of these fragments is then linked to detected words, which could be objects, actions, or attributes identified within the visual content. Subsequently, the generated fragments are reassembled into a sentence using a predefined

language template. Let us underline that the quality of the captioning process heavily relies on the sentence templates, with sentences always being produced with a syntactical structure.

In the pioneering work introduced in [224], a concept hierarchy of actions is constructed for the natural language description of human activities. Here, a Conditional Random Field (CRF) is used to establish a semantic representation for each video, by employing a template model for sentence generation. Additionally, [225] authors put forward a two-step method involving a Highest Vision Confidence (HVC) model and a Factor Graph Model (FGM). This process initially establishes confidences on the subject, verb, object, and scene elements. Subsequently, a factor graph model is implemented to deduce the most likely Subject-Verb-Object (SVO) triplet in the video. Finally, a sentence is constructed based on the considered template.

Although template-based language models can generate complete sentences, the descriptions produced are typically inflexible. Moreover, evaluation is often confined to a narrow domain with a restricted vocabulary. For any adequately rich domain, the complexity of rules and templates needed makes the manual design of templates impractical or overly costly.

8.3.2 Deep-learning based approaches

More recently, video captioning has been reformulated as a machine translation task [225], [226], leading to the development of the encoder-decoder paradigm (Figure 8.3) that is commonly used today. Within this framework, the encoder processes a set of video features and accumulates its hidden states. The resulting output state is then passed to a decoder, which generates a natural language caption based on the encoded information. Such an approach makes it possible to model complex video features, and thus generate captions that are more semantically meaningful than those obtained by rule-based methods. Moreover, the encoder-decoder paradigm can be trained in an end-to-end fashion, allowing the simultaneous optimization of both encoder and decoder. This leads to improved performances when applied on the video captioning task. We can identify two families of approaches that exploit the encoder-decoder paradigm: the visual-based approaches and the multimodal approaches.

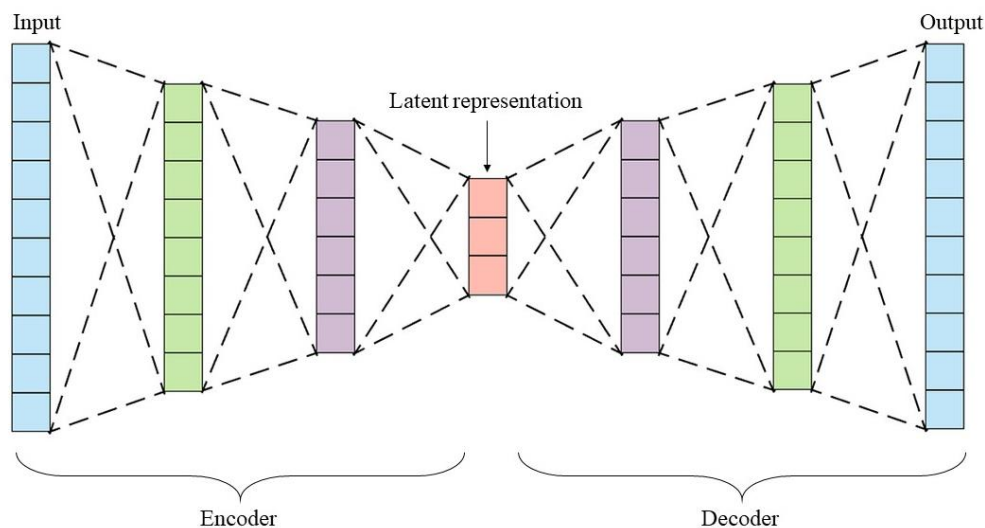


Figure 8.3. The paradigm of the encoder-decoder architecture.

8.3.2.1 Visual-based approaches

Visual-based approaches in video captioning focus primarily on extracting relevant visual information from video frames. Such approaches leverage computer vision techniques to analyze the visual content of the video and identify important elements such as objects, scenes and actions, together with their corresponding spatial and temporal relationships. In early works, the visual encoder is implemented as a 2D CNN applied to video frames. Thus, Venugopalan *et al.* [227] propose a framework where CNN features from each frame are averaged and provided as input to the decoder at every time step. Zhang *et al.* [228] introduce the GMNet model, incorporating a guidance module within the encoder-decoder model for video caption generation. GMNet facilitates word generation by considering both preceding and subsequent words in the caption. The model utilizes a soft attention mechanism and leverages InceptionV4 [229] to extract semantic features from the video.

To capture temporal dynamics within the video, the 2DCNN architecture has been later extended to 3DCNNs [69]. Xu *et al.* [230] introduced a two-module model for video captioning: the 'Proposal Module' that extracts features using 3D convolutional layers (C3D), and the Segment Proposal Network (SPN) for obtaining temporal segments. The model maps visual representation solely from the video to a common vector space, while the syntactic representation relies on the Part-of-Speech (POS) tagging structures of the video description. Hemalatha and Sekhar [231] introduce a video captioning approach that incorporates domain-specific decoders through the use of a domain classifier. The model utilizes ResNet152 for extracting 2D-CNN features and a 3D-CNN for extracting temporal features. To obtain a video representation, both the 2D-CNN and 3D-CNN features are processed using VLAD [232].

For sentence generation, many existing approaches rely on recurrent neural networks (RNNs) such as LSTM [233] and GRU [234] to generate the caption. Yao *et al.* [83], Donahue *et al.* [66] and Venugopalan *et al.* [227] use the LSTM architecture for yielding variable-length video descriptions. Guo *et al.* [235] further incorporate attention mechanisms within the LSTM model to refine the captions. Similarly, Zhang *et al.* [236] introduce a hierarchical decoder with temporal or spatial attention. The model implements a teacher-recommended learning system to leverage external language models and incorporate linguistic information.

Overall, visual-based approaches primarily focus on leveraging visual cues to generate accurate and descriptive captions. They are particularly effective in scenarios where the video content is predominantly visual and lacks significant audio or textual cues. However, in many applications videos represent a rich source of information, as they often contain multiple modalities such as visual, audio, and textual information (e.g., subtitles). Such modalities contribute to the overall meaning of the video and must be jointly considered to generate meaningful captions [237]. The multimodal approaches have gained popularity in video captioning task as they provide a more comprehensive understanding of the video.

8.3.2.2 Multimodal approaches

Currently, various methods adopt multimodal learning in video captioning tasks. Hessel *et al.* [237] use both automatic speech recognition (ASR) and video features to perform video captioning and claim that most of the enhancement in performance is attributable to the use of ASR. Similarly, Shi *et al.* [238] train their video captioning model on both visual and ASR inputs and demonstrate the benefits of adding the textual input to the overall understanding of the video. Inspired by such results, we have also considered in our work both visual and textual modalities.

However, multimodal video captioning also presents several challenges. One major one concerns the alignment between different modalities, as the timing and content of the visual and textual inputs may not always be perfectly synchronized [194]. Furthermore, the size and complexity of multimodal datasets can raise challenges for training models that are both accurate and efficient [109]. To tackle

such issues, several works [100], [194] use instructional videos [137], where the video and subtitles tracks are specifically aligned for video captioning tasks. While such videos are useful for training, they have specific structure and format that may not be representative of real-world scenarios [182]. This makes it difficult to generalize the model on unseen data.

Furthermore, real-world speech tends to be less structured, with key actions or events in the video not always corresponding to the same segments in the input transcript. To address the visually misaligned narrations, various approaches have employed contrastive learning between video and transcript. For instance, MIL-NCE [194] leverages weak and noisy training signals in instructional videos by combining multiple instance learning with contrastive learning. Meanwhile, VideoCLIP [239] constructs temporally overlapped pairs of video and text clips of varying lengths, aiming to enhance the quality and quantity of the pre-training dataset.

Traditionally, most existing methods have applied the contrastive loss to the outputs of visual and text encoders, typically before the multimodal fusion stage [111]. The primary aim of this loss is to establish alignment between the video and transcript during the pre-training phase. In our approach, we tackle the alignment challenge differently by incorporating the modality attention module. This module is specifically designed to bridge the gap between video and text modalities before feeding them into a transformer encoder. By exploiting cross-correlation, the modality attention module generates text-conditioned visual features and video-conditioned textual features, facilitating a more effective alignment. Conversely, the contrastive loss serves the purpose of aligning the multimodal representation of the input with its corresponding caption. In contrast to previous state-of-the-art models, we apply this loss to the output of the multimodal transformer.

8.3.2.2.1 Architecture

In view of the success of transformers in several domains, recent methods use this architecture as both encoder and decoder to tackle video captioning. Under this framework, three main training paradigms are encountered in the literature.

The *share-type* paradigm, illustrated in Figure 8.4 (a), includes Unicoder-VL [97], VL-BERT [98], UNITER [99], VideoBERT [100], and VideoAsMT [101]. In this case, the textual and visual modalities are fed into a single encoder that generates a unified representation. While computationally efficient, this approach suffers from modality entanglement due to the vast differences among various modalities [240]. This challenge stems from the fact that several modalities may interfere, particularly when there are numerous modalities and tasks involved [241]. It is challenging for a foundational model with a single module to strike a balance between the advantages of modality collaboration and the impact of modality entanglement across various modalities.

The *cross-type* paradigm, illustrated in Figure 8.4 (b), includes models like ViLBERT [102] and LXMERT [103]. Within this framework, multiple separate encoders are used to accommodate to the different interactions between modalities. In contrast to the single-stream input in the share-type, the two-stream input allows for interactions between different modalities at various representation depths. The cross-type approach can be more computationally demanding due the use of several cross-encoders.

Finally, the *joint-type* paradigm, illustrated in Figure 8.4 (c), is used by models such as SwinBert [242], UniVL [243] and MV-GPT [244]. This paradigm utilizes a two-stream input, similar to the cross-type architecture, allowing for effective capture of intra-modal features. However, in contrast to the cross-type, the joint-type architecture incorporates a single encoder to capture inter-modal dependencies. This approach strikes a balance between computational efficiency and the capacity to capture modality-specific features and interactions. For this reason, in our work we have also adopted the joint-type encoding paradigm.

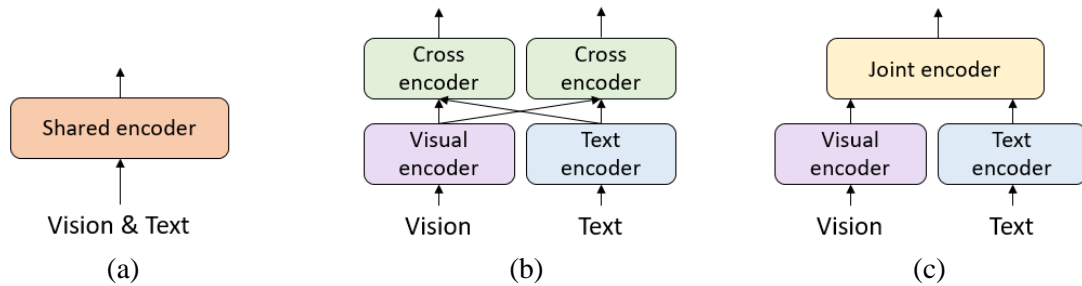


Figure 8.4. Various paradigms for video-text training. (a) Share-type; (b) Cross-type; and (c) Joint-type.

There are several architectures that can be considered for the decoder component in video captioning models. One common approach is to use RNNs, which generate the caption word by word, in a sequential manner. This method has the advantage of being able to capture long-term dependencies between words, but can suffer from slow convergence and difficulty in modeling complex relationships between video and language [4]. More recently, several studies [100], [242]–[244] have explored transformer-based models for video captioning, which show promising results due to their ability to capture long-range dependencies and relationships between different modalities. We follow this line of work and use the attention mechanism to sequentially generate the caption. We use both the encoder hidden states and the previously generated words in the caption as supervisory signals for the attention.

8.3.2.2.2 Training strategies

In recent years, vision-language pre-training has gained considerable popularity within the research community [245]–[248]. This approach involves an initial phase where multimodal models are pre-trained on extensive datasets in an unsupervised manner, followed by subsequent fine-tuning for specific downstream tasks (e.g. video captioning, action recognition, video question answering). Typically, the considered datasets comprise videos along with their associated transcripts, a resource that is abundantly available. These methods learn multimodal representations by formulating proxy tasks such as masked language modeling [243], [244], or vision-language matching [245], [246].

The paradigm of pre-training followed by fine-tuning for multimodal models is undeniably effective and has yielded remarkable results across various applications [111], [249]. However, it is essential to acknowledge that this approach comes with substantial resource requirements, primarily in terms of hardware, rendering it unfeasible for small-scale setups. This is particularly the case when considering multimodal models with billions of parameters, such as the GIT model [250], which has over 5 billion parameters and is pre-trained on 10.5 billion samples. Additional statistics for similar models can be found in section 8.6.4. The resource-intensive demands penalize the adoption and deployment of such approaches in the case of applications where the computational resources are limited/constrained, most often for economical reasons.

Within this context, let us note that pre-training undoubtedly enhances the model’s performances. Thus, comparing pre-trained models with models learnt from scratch is not entirely equitable. In our case, due to hardware constraints, we opt for an alternative strategy by forgoing pre-training altogether. Despite this, we demonstrate that competitive results can still be achieved. The proposed approach leverages the available resources efficiently, focusing on task-specific training without the need for massive pre-training datasets or extensive computational power. This resource-aware approach not only makes multimodal modeling accessible to a wider range of users and applications but also highlights the potential for effective multimodal model development in resource-constrained environments.

8.3.3 Evaluation metrics

Human evaluation of model performance, while highly valuable, poses challenges due to its resource-intensive nature, time consumption, and potential subjectivity. Consequently, the deployment of automatic, objective evaluation tools is necessary. The validity of such evaluation metrics is correlated to their alignment with human judgement. An effective evaluation metric should deliver reliable results despite potential linguistic alterations in the text. Such alterations could include synonym substitution, the addition of redundant words, modification of word sequence, or the abbreviation of sentences, provided the semantic integrity of the original content remains unaltered. In this section, we review the most common evaluation metrics used in the literature including BLEU, ROUGE, and METEOR.

8.3.3.1 BiLingual Evaluation Understudy (BLEU)

The BLEU [151] metric quantifies the quality of machine-generated translations by determining the proportion of n -grams in the machine translation that overlap with human-generated reference translations. The BLEU[B@ n] score is calculated using the following formula:

$$BLEU[B@n] = BF \times \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (8.1)$$

where BF denotes the brevity penalty factor, p_n signifies the geometric mean of the modified n -gram precision up to length N , and w_n represents the weight of the n -gram precision, with the sum of w_n terms normalized to 1. Let c_l denote the length of the machine translation, and r_l the length of the reference translation. The brevity penalty factor, BF , in Eq.(8.1) is computed as follows:

$$BF = \begin{cases} 1 & \text{if } c_l > r_l \\ e^{(1-\frac{r_l}{c_l})} & \text{if } c_l \leq r_l \end{cases} \quad (8.2)$$

The BLEU[B@ n] metric fundamentally relies on precision, without incorporating recall. This metric distinguishes words with synonymous meanings as separate entities, thereby penalizing even small variations in words. This limitation is mitigated by the METEOR metric, which tends to align more closely with human judgement due to its ability to recognize synonyms and consider recall in addition to precision.

8.3.3.2 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

The ROUGE-L [152](Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) metric assesses the quality of a generated summary by comparing it to reference summaries. This comparison involves measuring the overlap of n -grams and word pairs between the machine-generated summary and the human-crafted reference summaries.

The ROUGE-N variant calculates an n -gram recall between a system-generated summary and a collection of human-generated summaries, as described in the following equation:

$$ROUGE - N = \frac{\sum_{s \in \text{Reference summaries}} \sum_{s \in n\text{-gram}} \text{Count}_{\text{match}}(n - \text{gram})}{\sum_{s \in n\text{-gram}} \text{Count}(n - \text{gram})} \quad (8.3)$$

where n indicates the length of the n -gram, $\text{Count}(n - \text{gram})$, and $\text{Count}_{\text{match}}(n - \text{gram})$ respectively represent the maximum number of n -gram overlaps in a candidate summary and a set of reference summaries.

The ROUGE metric is primarily recall-based, given that the denominator of the equation represents the total sum of the number of n -grams present in the reference summary. In Eq.(8.3) if more reference

summaries are included into the metric, the number of matching n -grams in the denominator increases. Each addition to the reference pool broadens the distinct summaries' space. ROUGE-N evaluates various facets of text summarization by modulating the type of reference added to the pool. The numerator of Eq.(8.3) aggregates all the reference summaries, granting more weight to matching n -grams that appear in multiple references. Thus, ROUGE-N rewards machine-translated summaries that share a greater number of words with the reference summaries.

There exist several variations of the ROUGE metric. The ROUGE-L version uses a Longest Common Subsequence (LCS)-based F-Measure to establish the correlation between two text summaries A and B (where A is a reference summary and B is a model-generated summary). It calculates R_{LCS} , P_{LCS} , and F_{LCS} using the following formulas:

$$R_{LCS} = \frac{A + B}{a} \quad (8.4)$$

$$P_{LCS} = \frac{A + B}{b} \quad (8.5)$$

$$F_{LCS} = \frac{1 + \gamma^2 R_{LCS} P_{LCS}}{R_{LCS} + \gamma^2 P_{LCS}} \quad (8.6)$$

where $LCS(A,B)$ denotes the length of the LCS of summaries A and B, and $\gamma = \frac{P_{LCS}}{R_{LCS}}$.

8.3.3.3 Metric for Evaluation of Translation with Explicit Ordering (METEOR)

The METEOR [251] evaluation metric incorporates both precision and recall. Precision, denoted as P , is computed as the ratio of the number of unigrams in the machine-generated translation that correspond with those in the human-generated translation, over the total quantity of unigrams in the machine-generated translation. Recall, denoted as R , is determined by the proportion of unigrams in the machine-generated translation that overlap with those in the reference translation, over the aggregate amount of unigrams in the reference translation.

The harmonic mean of precision and recall, H_m , is calculated as follows:

$$H_m = \frac{10PR}{R + 9P} \quad (8.7)$$

The METEOR penalty is computed as follows:

$$Penalty = 0.5 \times \left(\frac{\#chunks}{\#unigrams\ matched} \right)^3 \quad (8.8)$$

The final METEOR score, M_s , is derived from the harmonic mean and the penalty, as expressed in the following equation:

$$M_s = H_m \times (1 - Penalty) \quad (8.9)$$

This score reflects the intersection of precision and recall, thus offering a more balanced assessment of the translation quality, addressing both false positives (captured by precision) and false negatives (captured by recall).

8.4 Proposed video captioning architecture

Figure 8.5 illustrates the synoptic scheme of the proposed approach, which comprises three fundamental elements: (1) the modality attention module, (2) the joint encoder and (3) the decoder. As a preprocessing step, we start by extracting the visual and textual embeddings.

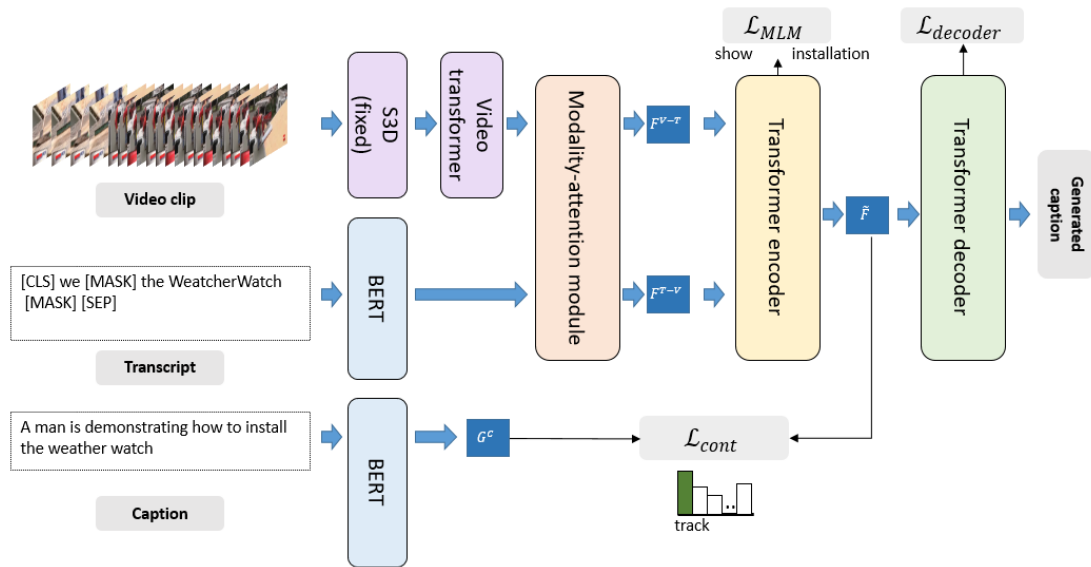


Figure 8.5. Overview of the proposed multi-modal architecture.

8.4.1 Feature extraction

The feature extraction process concerns the two components involved, which are the visual and textual (with both transcript and caption) data.

8.4.1.1 Visual feature representation

In order to acquire the visual representations, we utilize a uniform sampling approach to divide the video into N fixed length, non-overlapping clips of 16 frames each. The clips are then processed with the help of the S3D network [69], which is designed to learn robust video representations. Prior to use, the S3D model has been pre-trained on HowTo100M [137] with the MIL-NCE technique [194]. The feature activations before the final fully connected layer are extracted and we apply average pooling to generate a $d_v=1024$ -dimensional vector (the v subscript stands here for visual). Subsequently, a feed forward network that includes a linear projection, followed by the GeLU [195] activation function and layer normalization, is used to yield the final feature vector (of the same size d_v). The resulting visual features are represented as a $N \times d_v$ matrix, denoted by V . Let us underline that the S3D model is used only as a backbone for feature extraction and its weights are subsequently frozen.

A video transformer is further employed to effectively capture the dependencies between frames in video clips and learn the inherent temporal dynamics of video objects, actions, and scenes. This approach enables us to learn grounded visual features that are specifically optimized for the task of video captioning, without being restricted to pre-extracted features from external models. In addition, using a pre-extracted feature-based model with a transformer architecture can significantly reduce the

computational cost of training, as the S3D model can be pre-trained on large-scale video datasets and the transformer can be fine-tuned on a smaller dataset dedicated to video captioning.

To take into account the dynamic dependencies between clips, we employ temporal attention on the feature vector V . Our approach is motivated by the observation that video data often contains redundant information, and only a limited number of clips contain discriminative information that is relevant for the video captioning task. For this reason, a multi-head temporal attention mechanism is applied on the visual descriptor V . For each attention head $h \in \{1, \dots, H^v\}$ (where H^v denotes the number of visual attention heads), we first compute the associated $Query_h^v$, Key_h^v and $Value_h^v$ components defined as:

$$Query_h^v = VW_{query,h}^v; Key_h^v = VW_{key,h}^v \quad (8.10)$$

$$Value_h^v = VW_{value,h}^v \quad (8.11)$$

where $W_{query,h}^v$, $W_{key,h}^v$ and $W_{value,h}^v$ are three learnable matrices of size $(d_v \times \frac{d_v}{H_v})$.

The visual temporal attention for a given attention head is computed as:

$$Attention_h^v = softmax(\frac{Query_h^v(Key_h^v)^T}{\sqrt{d_v/H_v}})Value_h^v \quad (8.12)$$

where superscript T denotes the matrix transpose operator.

The attention heads are then concatenated under the form of a $(N \times d_v)$ matrix, denoted by $Attention^v$ and globally gathering the visual representation. An additional projection is considered in order to obtain the final visual representation, denoted by F^v and defined as:

$$F^v = Attention^v W_{convert}^v \quad (8.13)$$

where $W_{convert}^v$ is a learnable matrix of size $(d_v \times d)$. This final operation performs the dimensionality conversion of the visual feature to a common dimension d that will also be used for the textual representation.

8.4.1.2 Textual feature representation

Concerning the textual data, we consider the audio transcript (if the audio channel includes speech) as well as the video captions.

To obtain the audio transcript from the input video, we utilize the Whisper model [196] which is an ASR algorithm that exhibits human-level robustness in English speech recognition, even in the presence of background noise and reverberation.

Whatever the source (audio transcript or caption), the textual data undergo a tokenization process using WordPieces [197], which segments the text into sub-words using a vocabulary of $S_{voc} = 30,000$ tokens. The tokenized sequences are fed into the BERT-based uncased model [116], which performs the embedding. As recommended in [116], the first token in the input sequence is represented as a dedicated [CLS] token, and the final one is represented by a so-called [SEP] token. To achieve equal length for all the tokenized text sequences, we expand the sentence using padding, with the help of the [PAD] token. Let us denote by M the length of the padded tokenized sequences, which correspond to the maximal number of tokens that are allowed to appear in a given sentence. Let us also mention that a random masking of the tokens can also be considered. In this case, the input token is replaced by a dedicated token, denoted by [MASK].

The BERT approach also employs a self-attention mechanism, yielding in output a $(M \times d_{BERT})$ feature matrix, with $d_{BERT} = 768$, corresponding to the activations of the last BERT layer.

The text embedding approach is applied on both transcript and caption data. Similarly to the visual component, the transcript feature matrix, is finally converted into a $(M \times d)$ matrix denoted by F^t , with d being the common dimension considered also for the visual representation. The caption feature matrix, denoted by F^c , does not require projection onto a space of common dimension (see its utilization in section 8.5.2) and thus remains of size $(M \times d_{BERT})$.

Let us finally note that BERT encoder is fine-tuned separately for the transcript and the caption data.

8.4.2 Modality Attention module

Modeling visual and textual dynamics within and across modalities is a highly intricate task. To overcome such a challenge, we have developed a modality-attention module (Figure 8.6) that effectively captures both intra- and inter-modal relationships between the visual and audio transcript modalities. It is designed to bridge the gap between features F^v and F^t , which are generated from separate models trained on different tasks.

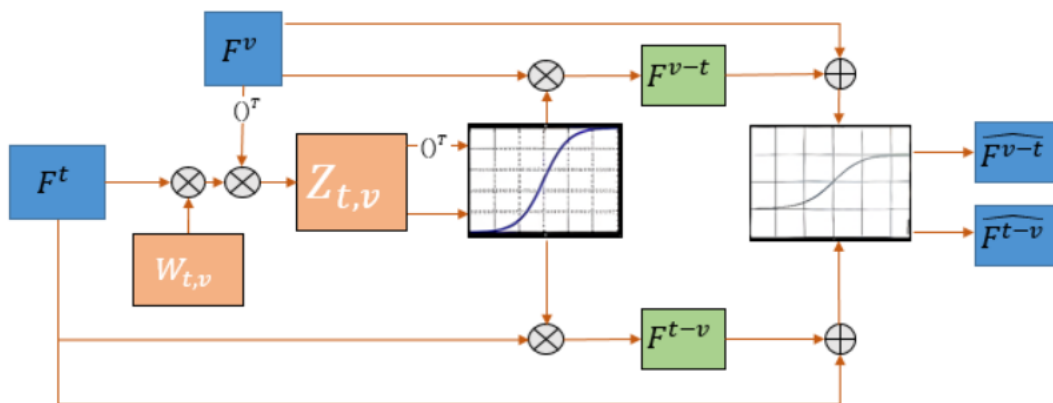


Figure 8.6. The modality attention module.

As we use real-life videos as input, in a majority of cases the feature vectors are not well-aligned. Figure 8.7 illustrates some video examples with their respective transcripts. We observe that people tend to speak in a disorganized manner, and the key actions or events in the video do not necessarily correspond to the same segment of the input text.



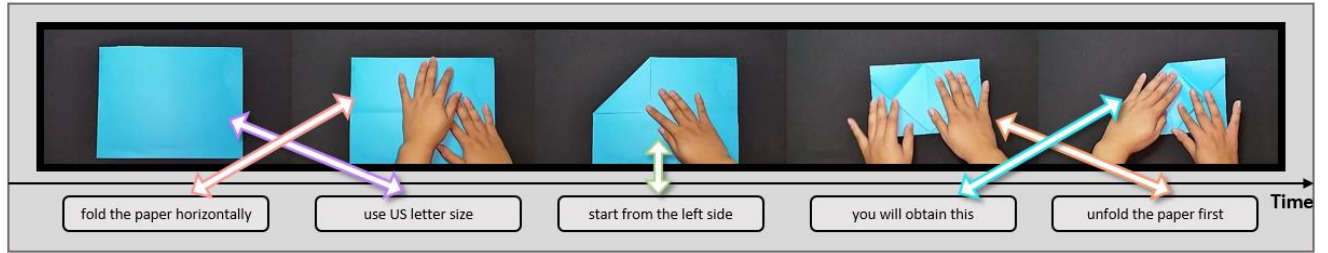


Figure 8.7. Video samples from MSRVTT dataset for which the transcript and video data are not well-aligned.

The objective is to create an embedding space that makes semantically-related visual-textual pairs of features appear closer together than unrelated pairs. This will enhance the alignment between F^v and F^t , and enable better modeling of the interactions between visual and audio transcript data. To this purpose, we consider the cross-correlation matrix $Z_{t,v}$:

$$Z_{t,v} = F^t W_{t,v} F^{vT} \quad (8.14)$$

where $W_{t,v}$ is a $(d \times d)$ learnable matrix and T denotes the transpose operator.

A high coefficient in the correlation matrix $Z_{t,v}$ indicates a strong relationship between the corresponding visual and textual features. To create cross-correlation visual-transcript (resp. transcript-visual) weights, we apply column-wise softmax over $Z_{t,v}$ (resp. $Z_{t,v}^T$), as described in the following equations:

$$F^{t-v} = F^t \text{softmax}(Z_{t,v}) \quad (8.15)$$

$$F^{v-t} = F^v \text{softmax}(Z_{t,v}^T) \quad (8.16)$$

This approach enables us to develop more distinctive and mutually constrained modality representations.

To avoid information loss during the cross-correlation phase, we have considered a dense skip connection technique. This means that we add the reweighted features F^{t-v} and F^{v-t} to the original representation of each modality, and regularize the result with the help of a \tanh function:

$$\hat{F}^{t-v} = \tanh(F^{t-v} + F^t) \quad (8.17)$$

$$\hat{F}^{v-t} = \tanh(F^{v-t} + F^v) \quad (8.18)$$

The modality attention module addresses the alignment issue between modalities. In equation (8.14), the cross-correlation matrix encodes the relationships between video and text features learned by the model through the trainable parameter $W_{t,v}$. Applying softmax to the matrix $Z_{t,v}$ enhances the discriminative power of the features. The model assigns higher weights to visual features when they exhibit strong correlations with textual features, and vice versa. This process potentially improves alignment between modalities. Specifically, it makes it possible to capture and emphasize the most salient correspondences between textual and visual elements. The resulting outcome, described in equations (8.15) and (8.16), is used to reweight the input features based on their correlation with the other modality. Finally, the skip connection technique detailed in (8.17) and (8.18) enforces the preservation of modality-specific information while adding non-linearity to the model.

8.4.3 Transformer encoder

In order to make the video and text fully interact, we design a transformer-based encoder. The transcript and visual features are first concatenated into a single global descriptor $F = [\hat{F}^{t-v} \mid \hat{F}^{v-t}]$, which is a matrix of size $(M + N) \times d$. The transformer architecture does not include any recurrent connections, which means that the order of the input tokens (or of video clips for the visual component) is lost during the process. To overcome this limitation, a position embedding technique is integrated. It consist in a trainable look-up table, where the embedding of each position in the input sequence is learned during training. To this purpose, we have followed the approach suggested in [116], described in the following equation:

$$E_{pos} = W_{pos}(pos_0, \dots, pos_{M+N}) \quad (8.19)$$

where W_{pos} of size $(M + N) \times d$ is a lookup table, mapping the position index of each token pos_i onto its corresponding vector representation.

In addition, a modality embedding is integrated, in order to differentiate between the visual and textual modalities:

$$E_{mod} = W_{mod}(\underbrace{0, \dots, 0}_M, \underbrace{1, \dots, 1}_N) \quad (8.20)$$

where W_{mod} of size $2 \times d$ is a lookup table, mapping the type of each modality (text: 0; video: 1) onto a vector representation.

The input to the encoder is defined as the sum of all these three features:

$$\mathcal{F}_0 = F + E_{pos} + E_{mod} \quad (8.21)$$

Our encoder comprises a number of L^{enc} self-attention layers. Each layer l consists of Multi-Head Self-Attention (MSA), layer normalization (LN) and Feed Forward Network (FFN). The considered layers, for $l \in \{0, 1, \dots, L^{enc} - 1\}$, are recursively computed as illustrated in Figure 8.8 and as described formally in the following equations:

$$\mathcal{F}'_l = MSA(LN(\mathcal{F}_{l-1})) + \mathcal{F}_{l-1} \quad (8.22)$$

$$\mathcal{F}_l = FFN(LN(\mathcal{F}'_l)) + \mathcal{F}'_l \quad (8.23)$$

The FFN consists of two linear projections separated by a GELU non-linearity [195].

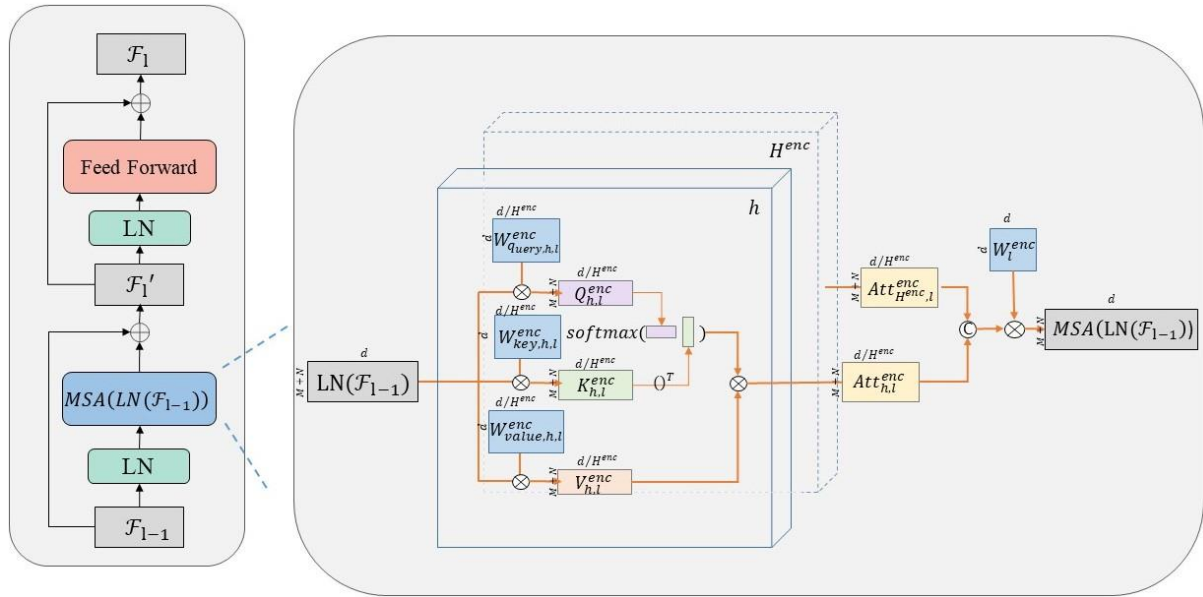


Figure 8.8. Overview of the encoder architecture. (left) Encoder block. (right) multi-head self-attention mechanism.

To enhance the model performance, we employ a multi-head attention mechanism, which splits the input into H^{enc} heads, allowing the model to attend to diverse parts of the input simultaneously. For each attention head $h \in \{1, \dots, H^{enc}\}$, we compute the attention sub-layers of the encoder as follows:

$$Att_{h,l}^{enc}(Q_{h,l}^{enc}, K_{h,l}^{enc}, V_{h,l}^{enc}) = \text{softmax}\left(\frac{Q_{h,l}^{enc} K_{h,l}^{encT}}{\sqrt{d/H^{enc}}}\right) V_{h,l}^{enc} \quad (8.24)$$

Here, the queries $Q_{h,l}^{enc} = LN(\mathcal{F}_l)W_{query,h,l}^{enc}$, keys $K_{h,l}^{enc} = LN(\mathcal{F}_l)W_{key,h,l}^{enc}$, and values $V_{h,l}^{enc} = LN(\mathcal{F}_l)W_{value,h,l}^{enc}$ represent linear projections of the multimodal input \mathcal{F}_l and d/H^{enc} is a scaling factor used to address the vanishing gradient issue.

Finally, the MSA is computed as follows:

$$MSA(LN(\mathcal{F}_l)) = \text{Concat}(Att_{1,l}^{enc}, \dots, Att_{H^{enc},l}^{enc})W_l^{enc} \quad (8.25)$$

where W_l^{enc} represents the learnable linear projection matrix.

The outputs of the various heads are concatenated and passed through a linear layer to obtain the final output $\mathcal{F}^{enc} = \mathcal{F}_{L^{enc}-1}^{enc}$ of size $(M+N) \times d$.

8.4.4 Transformer decoder

The objective of the decoder is to generate a caption $C = C(x_v, x_t)$ given the input video x_v and transcript x_t by maximizing the conditional probability $p(C|x_v, x_t)$. The caption C is represented as an ordered sequence of tokens $C = (c_1, c_2, \dots, c_{L_C})$. The joint probability can be recursively decomposed as follows:

$$p(C | x_v, x_t) = p(c_1 | x_v, x_t) \times p(c_2 | c_1, x_v, x_t) \times \dots \times p(c_{L_C} | c_{L_C-1}, \dots, c_1, x_v, x_t) \quad (8.26)$$

During training, the decoder generates one token at a time, conditionally to the previously generated tokens. However, by adopting such an approach, the errors can propagate and accumulate over time. In order to overcome this difficulty, we use the teacher-forcing technique [252] where the ground truth caption is forced to be provided until a certain token, selected in a random manner. Solely beyond this token, the model is allowed to generate its own ones. This technique stabilizes the training and limits the propagation of errors notably made in the early stages of decoding.

Formally, let $y^{c,n} = (t^{c,1}, \dots, t^{c,n})$ denote the sequence of decoded tokens up to token n . This sequence is iteratively providing new inputs $Y^{c,n}$ to the decoder, as described in the following equation:

$$\forall n \in \{1, 2, \dots, L_C\}, Y^{c,n} = dp(LN(emb(pad(y^{c,n})) + E_{pos}^c)) \quad (8.27)$$

where dp is the dropout layer, LN is the layer normalization, pad is the padding operator necessary to complete the $y^{c,n}$ sequence up to length L_C , emb is the embedding layer and E_{pos}^c is the positional embedding of the caption.

The transformer decoder consists in L^{dec} identical layers. Each layer l includes of a Masked-Multi-head Attention (MMA), layer normalization (LN), Multi-head Cross-Attention (MCA) and a Feed Forward Network.

The first layer is initialized as:

$$Y_0^c = (Y^{c,1}, \dots, Y^{c,L_C})$$

The subsequent layers, for $l \in \{1, \dots, L^{dec} - 1\}$, are recursively computed as illustrated in Figure 8.9 and as described formally in the following equations:

$$Y_l'^c = MMA(LN(Y_{l-1}^c)) + Y_{l-1}^c \quad (8.28)$$

$$Y_l''^c = MCA(LN(Y_l'^c), LN(\mathcal{F}^{enc})) + Y_l'^c \quad (8.29)$$

$$Y_l^c = FFN(LN(Y_l''^c)) + Y_l''^c \quad (8.30)$$

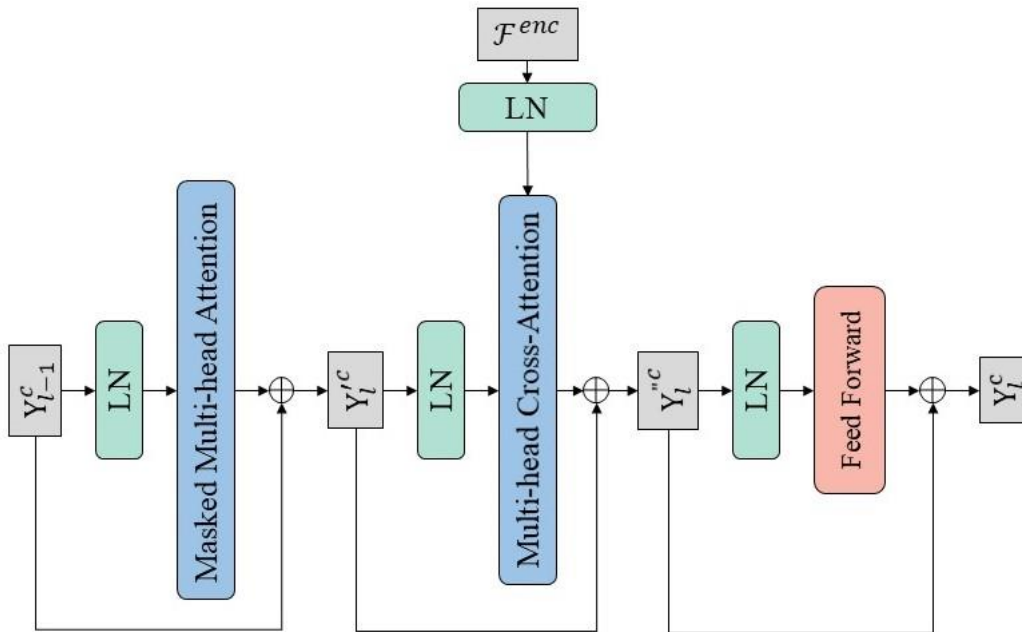


Figure 8.9. Overview of the decoder architecture.

The Masked Multi-head Attention mechanism represents a modification of the self-attention mechanism, consisting in a masking procedure whose goal is to prevent the decoder from attending future positions during training. This ensures the autoregressive property of the decoder, which is forced to get access solely to the tokens that precede the current position. The masking is achieved by setting the attention scores of future positions to a very large negative value. This ensures that the softmax operation applied to the attention scores assigns a probability close to zero to the future positions, thus effectively blocking their influence on the current position's representation. Formally, for each masked attention head $h \in \{1, \dots, H^{dec}\}$ and for each layer l , we compute the masked attention (*MAtt*) as:

$$MAtt_{h,l}^{dec}(Q_{h,l}^{MMA}, K_{h,l}^{MMA}, V_{h,l}^{MMA}) = \text{softmax}\left(\frac{Q_{h,l}^{MMA} K_{h,l}^{MMA^T}}{\sqrt{d/H^{dec}}} + \Lambda\right) V_{h,l}^{MMA} \quad (8.31)$$

where the queries $Q_{h,l}^{MMA} = LN(Y_l^c) W_{query,h,l}^{MMA}$, the $K_{h,l}^{MMA} = LN(Y_l^c) W_{key,h,l}^{MMA}$, the values $V_{h,l}^{MMA} = LN(Y_l^c) W_{value,h,l}^{MMA}$ represent linear projections of the decoder input Y_l^c . Here, Λ is the masking matrix of size $L_c \times L_c$. It is constructed such that the upper triangular portion (including the main diagonal) is filled with negative infinity values, and the lower triangular portion is filled with zeros.

We employ multi-head masked attention, and we concatenate the outputs of different heads as follows:

$$MMA(LN(Y_l^c)) = \text{Concat}(MAtt_{1,l}^{dec}, \dots, MAtt_{H^{dec},l}^{dec}) W_l^{MMA} \quad (8.32)$$

where W_l^{MMA} represents the learnable linear projection matrix.

Let us underline that during inference, the masked multi-head-attention is similar to the self-attention as the model does not have access to the future positions.

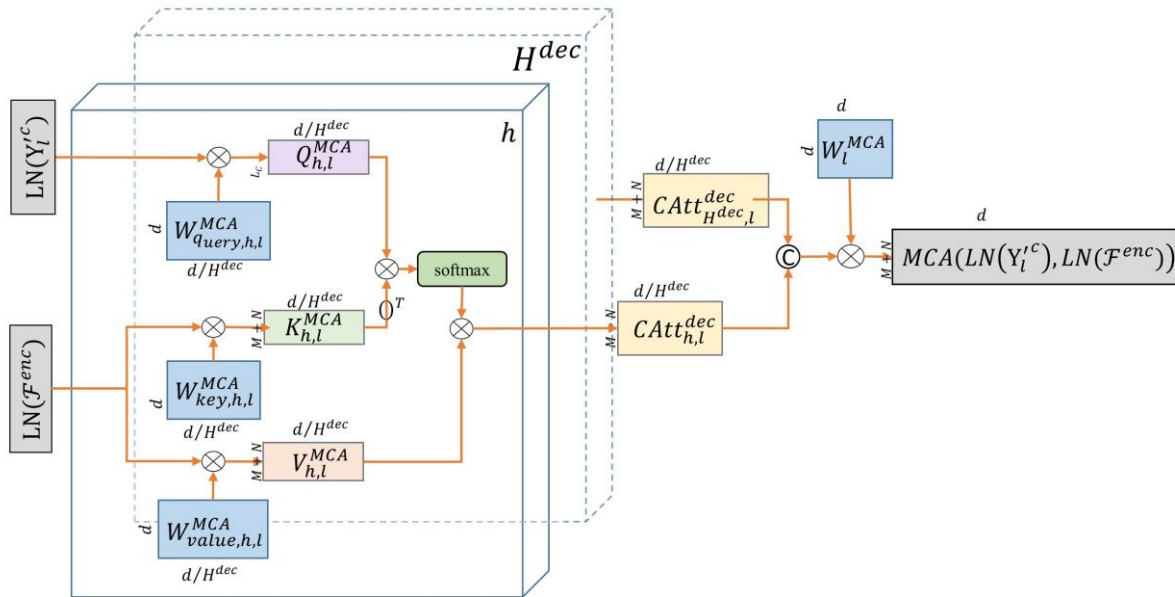


Figure 8.10. Multi-head Cross Attention process.

The second attention sub-layer is a Multi-Head Cross Attention (MCA), illustrated in Figure 8.10 and computed as follows:

$$Q_{h,l}^{MCA} = LN(Y_l^c) W_{query,h,l}^{MCA} ; \quad (8.33)$$

$$K_{h,l}^{MCA} = LN(\mathcal{F}^{enc}) W_{key,h,l}^{MCA} \quad (8.34)$$

$$V_{h,l}^{MCA} = LN(\mathcal{F}^{enc})W_{value,h,l}^{MCA} \quad (8.35)$$

where $W_{query,h,l}^{MCA}$, $W_{key,h,l}^{MCA}$, $W_{value,h,l}^{MCA}$ of size $d \times d/H^{dec}$ are learnable matrices. The cross attention $CAtt_{h,l}^{dec}$ is computed as follows:

$$CAtt_{h,l}^{dec}(Q_{h,l}^{MCA}, K_{h,l}^{MCA}, V_{h,l}^{MCA}) = softmax\left(\frac{Q_{h,l}^{MCA} K_{h,l}^{MCA T}}{\sqrt{d/H^{dec}}}\right)V_{h,l}^{MCA} \quad (8.36)$$

The outputs of the different cross attention heads are then concatenated and projected using a learnable matrix W_l^{MCA} as follows:

$$MCA(LN(Y_l^c), LN(\mathcal{F}^{enc})) = Concat(CAtt_{1,l}^{dec}, \dots, CAtt_{H^{dec},l}^{dec})W_l^{MCA} \quad (8.37)$$

The output of the final layer $Y_{L^{dec}}^c$ is used to determine the decoded token n as follows:

$$t^{c,n} = argmax(softmax(Y_{L^{dec}}^c W^{dec})) \quad (8.38)$$

using the learnable matrix W^{dec} and use the softmax function to compute the probability the token.

During inference, the model does not have access to the ground truth. Using the predicted output from previous time step can lead to a compounding error problem, where even small errors in the prediction can accumulate and result in poor performance. Therefore, we use the beam search decoding strategy to mitigate this problem. It is a heuristic algorithm that generates output sequences by keeping only the K most probable candidates at each step. Formally, at each time step n the decoder computes the probability distribution over the entire vocabulary for the next token as $p(c_n | c_{n-1}, \dots, c_1, x_v, x_t)$. Then we select the K candidates with the highest probabilities. For each candidate, the process is continued until an end token is generated or the maximum length is reached. Among all the generated candidates, the caption with the highest global probability is selected as output.

8.5 Training objectives

Three training objectives are considered to optimize the model: (1) masked language modeling, (2) contrastive learning and (3) caption generation.

8.5.1 Masked Language Modeling

Similar to BERT, we also randomly replace 15% of the tokens in the sentence with the special token [MASK], and then generate the masked tokens given the known tokens and video input. The Masked Language Modeling (MLM) loss function is defined as the cross-entropy loss between the predicted probability distribution over the vocabulary and the true distribution for each masked token:

$$\mathcal{L}_{MLM} = - \sum_{i=1}^{S_{mask}} \sum_{j=1}^{S_{voc}} y_{ij} \log(p_{ij}) \quad (8.39)$$

Here, S_{mask} is the number of masked tokens, S_{voc} is the size of the vocabulary, y_{ij} is the true probability of the j -th token for the i -th masked position and p_{ij} is the predicted probability of the j -th token for the i -th masked position.

8.5.2 Contrastive learning

Our goal is to create a system that can match a video x_v and transcript x_t to their correct caption C by calculating the dot product of their respective embeddings. We want to assign to incorrect captions a large distance, meaning that the dot product between their corresponding embeddings should be small.

Formally, we start by extracting the multimodal representation of the (video, transcript). We follow [188], and consider as a global representation of the multimodal input the embedding $\mathcal{F}^{enc}_{[CLS]}$ of the [CLS] token, which appears on the first position of the feature matrix $\mathcal{F}^{enc} = \{\mathcal{F}^{enc}_1 = \mathcal{F}^{enc}_{[CLS]}, \mathcal{F}^{enc}_2, \dots, \mathcal{F}^{enc}_{M+N}\}$. The global video-transcript representation is computed as:

$$\mathcal{F}_{global} = dp(\mathcal{F}_{[CLS]}W_{global} + b_{global}), \quad (8.40)$$

where W_{global} of size $d \times d$ and b_{global} of size d are learnt during training. We denote by $f(x_v, x_t)$ the function that associates a pair of video x_v and transcript x_t to their global representation \mathcal{F}_{global} .

Similarly, we extract the global representation of the caption embedding F_{CLS}^c (cf. Section 8.4.1.2) and project it as follows:

$$F_{global}^c = dp(F_{CLS}^c W_{global}^c + b_{global}^c), \quad (8.41)$$

where matrix W_{global}^c of size $d_{BERT} \times d$ and vector b_{global}^c of size d are learnt during training. Let us denote by $g(C)$ the function that associates the caption C to its global representation F_{global}^c .

The contrastive loss is then computed as:

$$\mathcal{L}_{Cont} = \max_{f,g} \sum_{i=1}^{batch_size} \log \left(\frac{e^{f(x_{v_i}, x_{t_i})^T \cdot g(c_i)}}{e^{f(x_{v_i}, x_{t_i})^T \cdot g(c_i)} + \sum_{(x_{v_j}, x_{c_j}, c_j) \in N_i} e^{f(v_j, t_j)^T \cdot g(c_j)}} \right) \quad (8.42)$$

Here, given a positive triplet of index i in the batch (x_{v_i}, x_{t_i}, c_i) of (video, transcript, caption), we construct the negative set N_i of negative triplet by concatenating incorrect captions c_j within the training batch to the (video, transcript) pair (x_{v_i}, x_{t_i}) as (x_{v_i}, x_{t_i}, c_j) with $c_j \neq c_i$.

8.5.3 Caption generation

The decoder loss measures the difference between the predicted caption and the ground truth caption using cross-entropy as follows:

$$\mathcal{L}_{decoder} = - \sum_{n=1}^{L_C} \log P(c_n | c_1, \dots, c_{n-1}, x_t, x_v) \quad (8.43)$$

The final loss function considered for our model is simply defined as the sum of all these three components:

$$\mathcal{L}_{model} = \mathcal{L}_{MLM} + \mathcal{L}_{Cont} + \mathcal{L}_{decoder} \quad (8.44)$$

8.6 Experiments and results

The experimental evaluation has been carried out on the publicly available dataset MSRVT [2], described in the following section.

8.6.1 Dataset

The MSRVT (Microsoft Research Video to Text) dataset is widely used for benchmarking video captioning methods. It spans over 20 domains, including sports, news, education, and how-to videos. The dataset comprises 10,000 video clips, with an average length of 20 seconds, and 200,000 natural language descriptions, which have been collected from crowd-workers, ensuring diverse and human-like language expressions. The videos have been crawled from YouTube, contributed by internet users, and thus correspond to real-life situations.

The MSRVT dataset raises several challenges, such as recognizing objects, actions, and scenes, as well as understanding the context and generating semantically meaningful captions. Additionally, it is worth noting that the MSRVT dataset comprises videos with both visual and audio modalities, which adds an extra level of complexity to the task of generating captions. Nevertheless, around 20% of the videos in the dataset have no audio channel, while others have non-English audio, making the task even more challenging with sparse modalities.

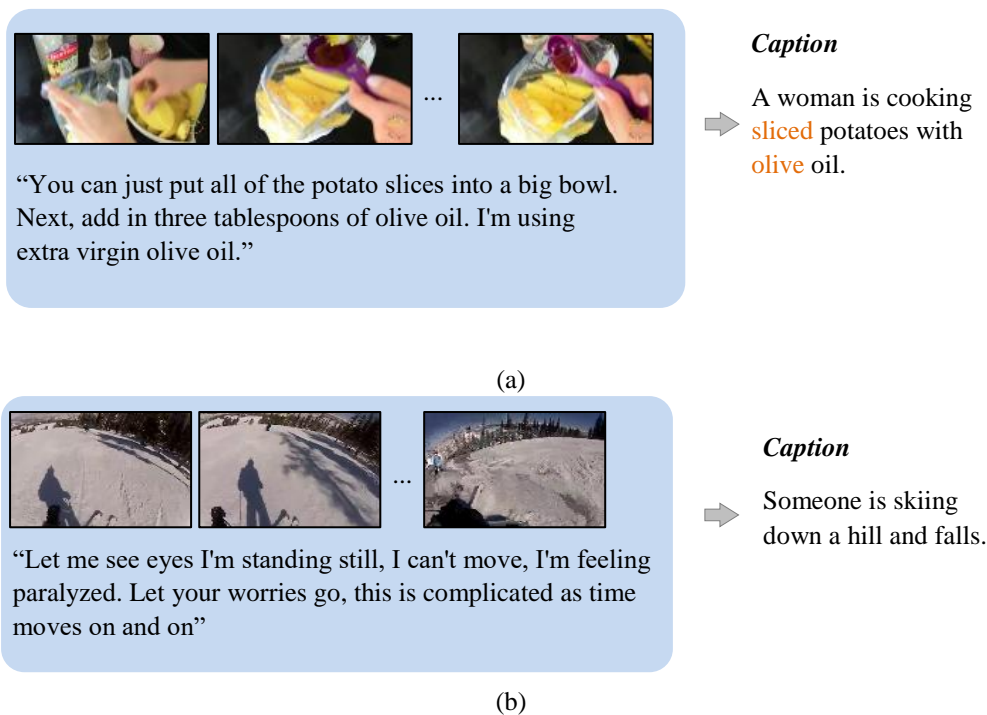


Figure 8.11. (a) Sample requiring both transcript and visual modalities for caption generation. (b) Sample requiring visual cues only.

In order to study the effect of each modality on the performance of the model, we have manually annotated two distinct subsets. The first subset, labeled as "vision and text" (534 samples), encompasses videos where both the visual features and transcript information contribute to the video captioning task. For example in Figure 8.11.(a), the transcript helps identifying specific ingredients such as oil type, difficult to discern solely from visual modality. The second subset, called the "vision only" (663 samples) subset, comprises videos where the task can be accomplished solely through visual cues. Some of these videos include silent or non-English speaking videos, where the transcript modality cannot be provided. Similarly, videos featuring sports or other activities that emphasize visual actions can be categorized in this subset. An example is illustrated in Figure 8.11.(b), the transcript represent the lyrics of music in the video and is not correlated to the caption. We study the performance of our model on these subsets to better understand the role of the transcript information in video captioning.

8.6.2 Implementation details

In the pre-processing stage, the videos are divided into $N = 48$ uniformly sampled clips. The clips are then processed with the help of the S3D model. Next, the transformer encoder is applied, with 6 layers to capture the sequential information in the 3D feature. Each block consist of $H_v = 12$ attention heads and a hidden size of $d_v = 1024$.

Regarding the transcript, we utilize the Whisper ASR model to extract the speech from the video. Our initial findings indicate that the quality of the ASR model has a notable influence on the overall performance. We apply the Whisper model on the entire video rather than on individual clips, as people commonly mention key objects or actions before or after they are shown in the video (Figure 8.7). We set the maximum number of tokens in a given phrase to $M=48$.

The model includes a 2-layer transformer encoder and a 3-layer transformer decoder, both consisting of 12 attention heads and a hidden size of $d = 768$. To accelerate the training process, we initialize the encoder and decoder weights with the pre-trained weights proposed by the model in [6]. The training process is conducted using 2 NVIDIA GeForce RTX 2080 GPUs over a period of 20 epochs, taking 4 days to complete. We use a linear learning rate schedule with a warm-up strategy, employing an initial learning rate of $1e-5$. To overcome the limited GPU memory, we use the gradient accumulation technique [252] with 16 steps in conjunction with a batch size of 256. This technique effectively increases the batch size and allows us to update the model's parameters with fewer samples, without sacrificing the accuracy of the gradient estimation. The final model is selected according to the best performance obtained on the validation set.

8.6.3 Ablation study

We have conducted an ablation study in order to determine the significance of each component within the framework. The study compares several combinations to evaluate their relative performances. More precisely, Table 8.1 reports the evaluation results obtained for the following methods:

- **text only**, which used only text as input, trained with the transformer encoder and decoder.
- **video only**, which used only video as input, also trained with the transformer encoder and decoder.
- **video-text**, which used both video and text as input but did not employ the modality-attention module.
- **MAM**, which adds the modality attention module (MAM) to the former.
- **MAM+init**, which uses the initialization of the encoder and decoder weights from the model in [6].
- **MAM+Cont**, which is trained with all objectives from scratch on MSRVT, including the contrastive loss with caption as input.
- **MAM+Cont+init**, which initializes the encoder and decoder weights using those of [6] and includes both modality-attention and contrastive loss objective techniques. We denote this complete architecture by **CapVT**.

Table 8.1. Ablation studies on MSRVT dataset

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE
Text only	0.6682	0.4684	0.33	0.2299	0.2087	0.4850
Video only	0.7643	0.6225	0.4878	0.3666	0.2637	0.5825
Video-text	0.7723	0.6456	0.5073	0.3782	0.2674	0.5881

MAM	0.7752	0.6507	0.5250	0.3972	0.2754	0.5987
MAM+Init	0.7909	0.6612	0.5320	0.4104	0.2803	0.6001
MAM+Cont	0.7979	0.6676	0.5373	0.4191	0.2893	0.6087
MAM+Cont+Init	0.8417	0.6784	0.5792	0.4408	0.3082	0.6291

The following evaluation metrics are retained to evaluate the performance of the models: BLEU (1-4) [151], METEOR [251] and ROUGE [152] (see 8.3.3). All scores range between 0 and 1, with higher values indicating better performances.

The ablation study results demonstrate that the complete CapVT model (MAM+Cont+init) outperforms other models with a BLEU4 score of 0.4408, indicating the importance of our training choices.

Pre-training the model on external large datasets can be beneficial, but this is often computationally expensive and requires significant hardware resources. To address this issue, we have used transfer-learning techniques to initialize the weights of our transformer encoder and decoder, which allowed us to leverage the knowledge learned from a larger dataset while reducing the computational load. This is observed with an improvement in performance of 1.2% and 2.17% for BLEU4 when comparing MAM to MAM+Init and MAM+Cont to CapVT, respectively. The study also highlights the importance of the contrastive loss objective in improving captioning quality, as removing it leads to a significant drop in performance (2.19% in terms of BLEU4). Additionally, incorporating textual information is crucial for generating accurate captions, as evidenced by the lower score achieved when only the video modality is considered. Finally, the results indicate that the visual modality is more informative than the textual one as we achieve better results when feeding only the visual modality as compared to feeding only the textual modality.

As part of our study, we also investigate how the quality of the generated captions is affected by different input modalities. For this purpose, we selected the first three baselines: text-only model, video-only model and video-text model. We have deliberately excluded the other models that employ additional strategies such as modality attention or contrastive loss. Our primary objective here is to solely examine the impact of the input modality on the model's performance.

Table 8.2. Performance comparison (BLEU4) across models using different input modalities on two subsets

Model	Vision-Only subset	Vision-text subset
Text-only model	0.1671	0.3247
Video-only model	0.3241	0.3051
Video-text model	0.3556	0.4192

We have assessed the performance of each baseline on videos that require only the visual modality to generate captions and those that require both visual and textual modalities. However, it was not feasible to label videos that require only textual modality in the MSRVT dataset as certain information, such as key objects/persons can only be perceived through visual cues and not through text. We have randomly selected 1179 test samples and manually labeled them as either "vision-only" (663 samples) or "vision and text" (534 samples) to evaluate the performance of each baseline model on these different

types of videos. Table 8.2 shows the performance comparison in terms of BLEU4 of the three models trained with different input modalities. The evaluation has been performed on the two subsets of samples that require either only visual modality or both visual and textual modalities to generate captions. The following conclusions can be drawn: (1) The effectiveness of video captioning models is heavily influenced by the input modalities. The different performances obtained on the two subsets underscore the significance of the dataset's modality composition. (2) The text-only model struggles to generate captions from visual content alone with a low score of 0.1671. (3) The video-only model performs well in a vision-only context, and may benefit from leveraging textual cues when available. Thus, adding the textual modality as input improves the performance with 11% on the vision-text subset. (4) The video-text model consistently outperforms the models relying on a single modality on the two subsets. This observation underscores the significance of multimodal approaches in video captioning. The ability to seamlessly integrate visual and textual information results in enhanced caption quality, making the model versatile and well-suited for real-world applications where both modalities are accessible.

8.6.4 Comparison with state of the art

To facilitate a meaningful comparison between our work and previous state-of-the-art models in the context of video captioning, we have examined key statistics pertaining to these models. Specifically, we have compiled comprehensive data encompassing model size (number of parameters), the scale of pre-training samples, hardware infrastructure employed (GPU/TPU), and the training duration. The detailed findings of this analysis are presented in Table 8.3, drawing from information extracted from the respective authors' publications and the survey introduced in [111].

Table 8.3. Statistics of video captioning models. PT stands for Pre-Training. x stands for unknown .

Method	Size	PT data scale	Hardware (GPUs/TPUs)	Training time
m-PLUG2 [245]	600M	766M	16 NVIDIA A100 GPUs	-
GIT [250]	681M	800M	x NVIDIA A100	-
GIT2 [250]	5.1B	10.5B	x NVIDIA A100	-
CLIP-DCD [248]	425M	400M	-	-
VAST [246]	1.3B	324M	64 Tesla V100	-
VideoCoca [247]	2.1B	144M	128 CloudTPU v4	6 hours
UniVL [243]	198M	136M	8 NVIDIA Tesla V100 GPUs	14 days
OA-BTG [226]	-	No PT	-	-
VideoAsMT [101]	286M	136M	-	-
SwinBert [242]	198M	No PT	Nvidia GPUs	V100 -

OpenBook [253]	-	No PT	-	-
CapVT	198M	No PT	2NVIDIA GeForce2080 GPUs	4 days

For a fair comparison, we have retained models that are comparable to ours, including OA-BTG [226], VideoAsMT [101], SwinBert [242], and OpenBook [253]. Additionally, we have retained the UniVL [243] model, as we leveraged its weights to initialize the encoder and decoder parameters in our own approach. This approach aims to deliver a comprehensive and equitable evaluation of our method in relation to its peers, thus establishing a clear understanding of its performance within a defined resource context.

Table 8.4 presents a comparison of CapVT with the retained models on the MSRVTT dataset. CapVT outperforms previous methods by a significant margin of 1.28%, 1.52%, and 1.21% in terms of BLEU4, METEOR, and ROUGE, respectively. Notably, even our model without encoder-decoder initialization (*MAM+Cont*) achieves comparable results, highlighting the effectiveness of modality fusion using modality attention and the importance of caption information in guiding the training. We anticipate that further improvements can be achieved by integrating a vision-language, end-to-end pre-training phase on the whole model. The results obtained demonstrate the pertinence of the CapVT model and its potential for achieving superior performance in video captioning tasks.

Table 8.4. Comparison with state of the art.

Method	BLEU4	METEOR	ROUGE
OA-BTG [226]	0.4140	0.2820	-
VideoAsMT [101]	0.417	0.285	-
UniVL [243]	0.4179	0.2894	0.6087
SwinBert [242]	0.419	0.299	0.621
OpenBook [253]	0.428	0.293	0.617
MAM+Cont	0.4191	0.2803	0.6087
CapVT	0.4408	0.3082	0.6291

8.6.5 Qualitative results

Some examples of results obtained on MSRVTT are illustrated in Figure 8.12, Figure 8.13 Figure 8.14. The results indicate that the quality of the predicted captions is affected by various factors, including the availability of audio and visual information, the complexity of the content and the accuracy of the ASR. When the transcripts are pertinent (with salient words represented in purple in Figure 8.12), combining

textual and visual modalities leads to a precise captioning. In contrast, in the absence of the audio (Figure 8.14) or more generally when the transcript channel is not coherent with the content (transcripts represented in red in Figure 8.13), the predictions rely only on visual clues and may not be as informative. In general, the accuracy of the predicted captions is largely influenced by the type and quality of the input data. In all cases, incorporating multimodal approaches can enhance the precision of the predictions.



ASR: When planning on going for a jog, be sure to lock the front swivel wheel and use the tether strap for maximum safety. A quick release trigger fold also makes this **stroller** easy to fold.

GT: A person is demonstrating the flexibility of a movement assistance device

Prediction: A woman is demonstrating how to use a **stroller**



ASR: What's up everybody, how are you all doing? Today I'm going to **teach my puppy how to sit down**. So let's check out how to teach my puppy, how to teach him to sit. Make sure you're gonna grab a good treat.

GT: A bald man with orange sunglasses tries to teach his puppy to sit down

Prediction: A man is talking about how to **teach a puppy how to sit down**



ASR: And now, on behalf of everyone involved in the research and development of the FPT, I want to **thank you for** your expressed confidence in our **product** and for allowing us to share in your efforts to **improve your feminine health**.

GT: A middle aged woman encourages you to use her product to increase your feminine health

Prediction: A middle aged woman **encourages you** to use her **product** to **increase your feminine health**.



ASR: One of the downsides of **Apple Pay** is that you really can't use it at a lot of places. Apple says that it has agreements with 220,000 store locations. That sounds like a big number, but it's really just 5.5% of all retail locations.

GT: A reporter in black coat discussing the down sides of apple pay

Prediction: A man in a suit is talking about **apple pay**

Figure 8.12. Qualitative results from MSRVTT dataset. Samples requiring both textual and visual modalities to generate the caption.



ASR: Um, so. About the what the hell is gonna happen here. I honestly don't know, but we're just gonna fly over here, figure it out! Blow up! Drop your bomb! I really did a good job there, I think

GT: Gameplay footage of someone playing a game

Prediction: A person is playing a video game



ASR: I bring it in when I got you some courtesy you gotta know You are so respectable You are so respectable You are so respectable

GT: There are some women dancing on the floor with music

Prediction: A group of people are dancing in a gym



ASR: I don't know if you can see the velveteros. I want to see what it looks like. We're coming. You're hungry. You're hungry. I'm hungry.

GT: A girl is playing the sims

Prediction: A person is playing a video game

Figure 8.13. Qualitative results from MSRVTT dataset. The ASR is not aligned with the content of the video.



ASR: \emptyset

GT: A girl and two men posing for a photo

Prediction: A still image of a man and a woman are shown



ASR: \emptyset

GT: A man is dicing an onion very quickly

Prediction: A person is slicing a red onion



ASR: \emptyset

GT: A 360 degree view of an Audi car

Prediction: A car is shown

Figure 8.14. Qualitative results from MSRVT dataset. Samples with no audio channel.

8.7 Conclusion

In this work, we have introduced CapVT, a novel architecture that efficiently exploits and combines rich information from both visual and transcript modalities for multimodal video captioning. The proposed modality-attention module and contrastive learning technique makes it possible to enhance the representation of inter-modal relationships, leading to a new state-of-the-art performance on the MSRVT dataset with respect to various evaluation metrics. The proposed model achieves a BLEU4 score of 0.4408, a METEOR score of 0.3082, and a ROUGE score of 0.6291 representing an improvement of 1.28%, 1.52%, and 1.21% respectively with respect to the state of the art. Our comprehensive study of each training strategy demonstrates the effectiveness of the CapVT model and its potential for achieving superior performance in the video captioning task. Furthermore, the study of

the effect of the input modalities involved highlights the effectiveness of our training strategies in improving the model's ability to generate accurate captions that rely on both text and visual information.

We have also found that the performance gain strongly depends on the nature of the data in different categories. This indicates a need for further research to develop more effective training methods that can take into account in a fine-grained manner the data characteristics of various categories. Future work could explore pre-training on larger datasets to further improve the performance of our approach. Large-scale pre-training allows the model to learn and capture the intricate correlations and interactions between different modalities. It facilitates the comprehension of complex multimodal patterns which may not be discernible in smaller, more constrained datasets. Another potential avenue can be the exploration of knowledge-augmented models. External knowledge sources can enhance the contextual understanding of video content, improve caption accuracy, and ensure domain relevance. They offer potential solutions to handle ambiguous or limited sensory cues, adapt to evolving content, and reduce biases. While knowledge-enhanced NLP models are widely studied, the exploration of knowledge-enhanced vision and multimodal models is a relatively uncharted territory, presenting an exciting opportunity for further research.

9 CONCLUSION AND PERSPECTIVES

9.1 Conclusion

This thesis represents a comprehensive exploration of the challenges faced by documentalists at France TV and the innovative application of deep learning techniques to enhance their workflows and improve the quality and consistency of their work. The research spans a wide spectrum of tasks, from unimodal models for landmark recognition, camera motion estimation, and scene identification to advanced multimodal models for tasks such as video question answering and video captioning.

Our first contribution in this research is centered around the development of a data-driven approach for classifying shots based on the type of camera motion. This problem is critical in the field of multimedia analysis, as it enables the automatic identification and categorization of shots in video content, which can have a profound impact on video production, content indexing, and user experience enhancement. To achieve this, we implemented a novel model based on 3D convolutional neural networks with residual blocks. These architectural choices were inspired by the success of similar techniques in action recognition. One of the key challenges we faced in this task was the scarcity of data specifically tailored to camera motion classification. To overcome this limitation, we employed a transfer-learning approach. We initiated our model's training on the Kinetics action recognition dataset, even though its content seemed unrelated to our immediate purpose. Interestingly, we found that the derived feature maps from this dataset contained essential spatio-temporal cues that could be exploited for our task. Moreover, our research introduced a semi-automatic method for dataset construction, which reduced the human intervention required for annotating and curating the data. Our model achieved an accuracy rate of 94% highlighting its effectiveness.

Our second contribution addresses the task Video Question Answering. One of the primary challenges in building a VideoQA model lies in the fundamental heterogeneity between the visual and textual modalities, as well as the inherent quadratic complexity of transformers. To overcome these challenges, we introduced a framework that integrates a lightweight transformer with a cross-modality module. The latter serves as a bridge between the visual and textual aspects of the task, facilitating the mutual learning of text-conditioned visual features and video-conditioned textual features. To test the robustness of our model, we introduced an adversarial testing scenario that involved rephrased questions, a reflection of the linguistic variations that occur in practical VideoQA applications. This test highlighted the adaptability of our model in handling diverse forms of input questions, showcasing its real-world applicability. The empirical evaluation of our model was conducted on the MSVD-QA and MSRVTT-QA benchmark datasets, achieving accuracy rates of 44.96% and 41.88%, respectively. These scores surpassed those of state-of-the-art methods and validated the efficacy of our proposed methodology.

Finally, we have tackled the challenge of video captioning. We introduced a novel framework known as CapVT, a multimodal architecture designed to capture intricate and meaningful relationships between visual and textual data. The framework incorporates a modality-attention module which allows the model to focus on the most relevant aspects of both visual and textual modalities. By emphasizing the salient features in each modality, the model can effectively align visual and textual information to generate more coherent and contextually relevant captions. We also incorporated contrastive learning to help the model distinguish and understand the relationships between different elements in the data. In the context of video captioning, this is invaluable for capturing the nuanced interdependencies between the video content and the corresponding textual descriptions. The effectiveness of the CapVT framework was thoroughly evaluated on the MSRVTT dataset. Our model achieves a strong performance across a range of evaluation metrics, including BLEU4, ROUGE, and METEOR, achieving scores of 0.4408, 0.6291, and 0.3082, respectively.

9.2 Future work

For future work, we outline several significant directions for research and development. To adapt multimodal models for industrial applications at France TV, one key area for future exploration is language-specific adaptation. While our experiments have been conducted using public datasets in English, extending the models to handle French content is essential for addressing the unique linguistic characteristics and demands of the French-speaking audience. Moreover, as we transition from experimental datasets to real-world TV programs, a critical consideration is the domain discrepancy. Future work can focus on bridging this gap to ensure that the models' performance is robust and reliable in the actual broadcast environment.

Another promising avenue for enhancing our model's performance and aligning it more closely with the unique requirements of documentalists, is the integration of Reinforcement Learning from Human Feedback (RLHF) techniques. By incorporating RLHF, we can create a feedback-driven model adaptation framework that leverages the valuable insights provided by documentalists. This approach not only empowers the model to continuously learn and improve but also ensures that it caters directly to the specific needs and expectations of the documentalists themselves. Hence, we can foster a dynamic feedback loop, enabling documentalists to actively participate in refining and optimizing the model's performance, ultimately resulting in a solution that is more finely attuned to their tasks and objectives.

Efficiency is another paramount concern, especially for industrial applications where real-time processing is crucial. While our research has introduced lightweight multimodal models, there is still room for optimization, particularly in reducing the inference time. Future research efforts may involve exploring advanced techniques such as model pruning and knowledge distillation to streamline the computational demands of these models, making them more suitable for deployment in industrial settings.

In the context of video captioning, our work has primarily concentrated on generating captions for short video formats. However, there is an emerging need for dense video captioning, which aims to produce text descriptions for a series of events in untrimmed videos. To achieve this, one potential avenue for future exploration involves adopting hierarchical structures in the models. These structures can better capture the long-term context and narrative flow within videos, ultimately leading to more informative and coherent captions. In addition, solutions that minimize the reliance on pre-segmentation, such as detecting and interpreting events within videos in a continuous and holistic manner, will be a critical area of study. Another potential avenue can be the exploration of knowledge-augmented models. External knowledge sources can enhance the contextual understanding of video content, improve caption accuracy, and ensure domain relevance. They offer potential solutions to handle ambiguous or limited sensory cues, adapt to evolving content, and reduce biases. While knowledge-enhanced NLP models are widely studied, the exploration of knowledge-enhanced vision and multimodal models is a relatively uncharted territory, presenting an exciting opportunity for further research.

In the context of video question answering, we have predominantly addressed factoid questions, as this field is still in its early stages of development. However, the significance of inference questions cannot be understated, as they delve into the nuanced relationships and connections within video content. Future research in this domain will focus on advancing the capabilities of models to answer inference questions, which may require sophisticated reasoning, contextual understanding, and a deeper analysis of the video's content. More broadly, there has been a growing interest in the development of universal and task-agnostic models. The research aims to create multimodal models capable of excelling across a broad spectrum of unimodal and multimodal downstream tasks, each with its unique characteristics and requirements. While significant strides have been made, there are several areas of future work that hold promise and merit exploration. This may involve a deeper exploration of the intricate interactions and synergy across different modalities, as well as the development of methods that allow fine-grained semantic alignments to naturally emerge. As the scale of multimodal training data expands, addressing

issues related to noise and data heterogeneity, training strategies, and model efficiency is becoming increasingly important. A comprehensive understanding of the strengths of Transformers, particularly their ability to encode implicit knowledge, aggregate non-local patterns, and handle domain gaps, will likely play a pivotal role in shaping the future of task-agnostic models.

List of Publications

- [1] K. Ouenniche, R. Tapu, and T. Zaharia, “A Deep Learning-Based Approach for Camera Motion Classification,” in *Proceedings - European Workshop on Visual Information Processing, EUVIP*, 2021. doi: 10.1109/EUVIP50544.2021.9483961.
- [2] K. Ouenniche, R. Tapu, and T. Zaharia, “Conditional Cross Correlation Network for Video Question Answering,” in *Proceedings - 17th IEEE International Conference on Semantic Computing, ICSC 2023*, 2023. doi: 10.1109/ICSC56153.2023.00011.
- [3] K. Ouenniche, R. Tapu, and T. Zaharia, “Vision-text cross-modal fusion for accurate video captioning,” *IEEE Access*, pp. 1–1, 2023, doi: 10.1109/ACCESS.2023.3324052.
- [4] K. Ouenniche, R. Tapu, and T. Zaharia, “Multi-modal video question-answering with cross-correlation and attention-based contextualization,” *International Journal of Semantic Computing (IJSC)*, submitted in April 2023.

REFERENCES

- [1] D. Xu *et al.*, “Video question answering via gradually refined attention over appearance and motion,” in *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, 2017. doi: 10.1145/3123266.3123427.
- [2] J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A Large Video Description Dataset for Bridging Video and Language,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 5288–5296. doi: 10.1109/CVPR.2016.571.
- [3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007. doi: 10.1109/CVPR.2007.383172.
- [4] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017.
- [5] K. Ozaki and S. Yokoo, “Large-scale Landmark Retrieval/Recognition under a Noisy and Diverse Dataset,” Jun. 2019, Accessed: Jun. 08, 2023. [Online]. Available: <https://arxiv.org/abs/1906.04087v2>
- [6] M. J. Swain and D. H. Ballard, “Color indexing,” *Int J Comput Vis*, vol. 7, no. 1, pp. 11–32, 1991, doi: 10.1007/BF00130487/METRICS.
- [7] A. Mojsilovic, J. Kovacevic, J. Hu, R. J. Safranek, and S. K. Ganapathy, “Matching and retrieval based on the vocabulary and grammar of color patterns,” *International Conference on Multimedia Computing and Systems -Proceedings*, vol. 1, pp. 189–194, 1999, doi: 10.1109/MMCS.1999.779145.
- [8] A. Laine and J. Fan, “Texture Classification by Wavelet Packet Signatures,” *IEEE Trans Pattern Anal Mach Intell*, vol. 15, no. 11, pp. 1186–1191, 1993, doi: 10.1109/34.244679.
- [9] D. G. Lowe, “Object recognition from local scale-invariant features,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, 1999, doi: 10.1109/ICCV.1999.790410.
- [10] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int J Comput Vis*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94/METRICS.
- [11] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image Vis Comput*, vol. 22, no. 10, pp. 761–767, Sep. 2004, doi: 10.1016/J.IMAVIS.2004.02.006.
- [12] K. Mikolajczyk *et al.*, “A Comparison of Affine Region Detectors,” *Int J Comput Vis*, doi: 10.1007/s11263-005-3848-x.
- [13] Y. Ke and R. Sukthankar, “PCA-SIFT: A more distinctive representation for local image descriptors,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, doi: 10.1109/CVPR.2004.1315206.

- [14] R. Arandjelovic and A. Zisserman, “Three things everyone should know to improve object retrieval,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2911–2918, 2012, doi: 10.1109/CVPR.2012.6248018.
- [15] H. Bay, T. Tuytelaars, L. Van Gool, A. Leonardis, H. Bischof, and A. Pinz, “SURF: Speeded up robust features,” *Lecture Notes in Computer Science*, vol. 3951, pp. 404–417, Jan. 2006, doi: 10.1007/11744023_32.
- [16] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 1470–1477, 2003, doi: 10.1109/ICCV.2003.1238663.
- [17] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the Fisher kernel for large-scale image classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6314 LNCS, no. PART 4, pp. 143–156, 2010, doi: 10.1007/978-3-642-15561-1_11/COVER.
- [18] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3304–3311, 2010, doi: 10.1109/CVPR.2010.5540039.
- [19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
- [20] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512–519, Sep. 2014, doi: 10.1109/CVPRW.2014.131.
- [21] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, “Neural codes for image retrieval,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014. doi: 10.1007/978-3-319-10590-1_38.
- [22] J. Y. H. Ng, F. Yang, and L. S. Davis, “Exploiting local features from deep networks for image retrieval,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2015. doi: 10.1109/CVPRW.2015.7301272.
- [23] G. Toliás, R. Sivic, and H. Jégou, “Particular object retrieval with integral max-pooling of CNN activations,” in *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016.
- [24] E. van der Spoel *et al.*, “Siamese Neural Networks for One-Shot Image Recognition,” *ICML - Deep Learning Workshop*, vol. 7, no. 11, 2015.
- [25] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, “From generic to specific deep representations for visual recognition,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2015. doi: 10.1109/CVPRW.2015.7301270.
- [26] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, “Visual Instance Retrieval with Deep Convolutional Networks,” *Kyokai Joho Imeji Zasshi/Journal of the Institute of Image Information and Television Engineers*, vol. 73, no. 5, 2019, doi: 10.3169/ITEJ.73.956.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.90.

- [28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008. doi: 10.1109/CVPR.2008.4587635.
- [29] H. Jegou, M. Douze, and C. Schmid, "Large scale image search," in *Proceedings of the 11th IAPR Conference on Machine Vision Applications, MVA 2009*, 2009.
- [30] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2 A large-scale benchmark for instance-level recognition and retrieval," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/CVPR42600.2020.00265.
- [31] "Flickr," <https://www.flickr.com/services/api/>.
- [32] D. Comaniciu and P. Meer, "Distribution free decomposition of multivariate data," *Pattern Analysis and Applications*, vol. 2, no. 1, 1999, doi: 10.1007/s100440050011.
- [33] K. Mikolajczyk and J. Matas, "Improving descriptors for fast tree matching by optimal linear projection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007. doi: 10.1109/ICCV.2007.4408871.
- [34] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016. doi: 10.1007/978-3-319-46604-0_48.
- [35] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marqués, and X. Giró-I-nieto, "Bags of local convolutional features for scalable instance search," in *ICMR 2016 - Proceedings of the 2016 ACM International Conference on Multimedia Retrieval*, 2016. doi: 10.1145/2911996.2912061.
- [36] R. Tapu, B. Mocanu, and T. Zaharia, "DEEP-AD: A Multimodal Temporal Video Segmentation Framework for Online Video Advertising," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2997949.
- [37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
- [38] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 40, no. 6, 2018, doi: 10.1109/TPAMI.2017.2723009.
- [39] "film-grab," <https://film-grab.com/>.
- [40] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," Oct. 2017.
- [41] J. G. Kim, Hyun Sung Chang, J. Kim, and H. M. Kim, "Efficient camera motion characterization for MPEG video indexing," in *IEEE International Conference on Multi-Media and Expo*, 2000. doi: 10.1109/icme.2000.871569.
- [42] W. J. Gillespie and D. T. Nguyen, "Robust estimation of camera motion in MPEG domain," in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 2004. doi: 10.1109/tencon.2004.1414440.

- [43] R. Tapu, B. Mocanu, A. Bursuc, and T. Zaharia, "A smartphone-based obstacle detection and classification system for assisting visually impaired people," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013. doi: 10.1109/ICCVW.2013.65.
- [44] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011. doi: 10.1109/ICCV.2011.6126544.
- [45] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry for ground vehicle applications," *J Field Robot*, vol. 23, no. 1, 2006, doi: 10.1002/rob.20103.
- [46] Y. Jiang, Y. Xu, and Y. Liu, "Performance evaluation of feature detection and matching in stereo visual odometry," *Neurocomputing*, vol. 120, 2013, doi: 10.1016/j.neucom.2012.06.055.
- [47] O. Naroditsky, X. S. Zhou, J. Gallier, S. I. Roumeliotis, and K. Daniilidis, "Two efficient solutions for visual odometry using directional correspondence," *IEEE Trans Pattern Anal Mach Intell*, vol. 34, no. 4, 2012, doi: 10.1109/TPAMI.2011.226.
- [48] H. E. Benseddik, O. Djekoune, and M. Belhocine, "SIFT and SURF Performance Evaluation for Mobile Robot-Monocular Visual Odometry," *Journal of Image and Graphics*, 2014, doi: 10.12720/joig.2.1.70-76.
- [49] A. Cumani, "Feature Localization Refinement for Improved Visual Odometry Accuracy," *International Journal of Circuits, Systems and Signal Processing*, vol. 5, no. 2, 2011.
- [50] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2008. doi: 10.1109/IROS.2008.4651147.
- [51] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," 2013. doi: 10.5244/c.2.23.
- [52] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans Pattern Anal Mach Intell*, vol. 32, no. 1, 2010, doi: 10.1109/TPAMI.2008.275.
- [53] M. A. Fischler and R. C. Bolles, "Random sample consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Commun ACM*, vol. 24, no. 6, 1981, doi: 10.1145/358669.358692.
- [54] N. Nourani-Vatani and P. V. K. Borges, "Correlation-based visual odometry for ground vehicles," *J Field Robot*, vol. 28, no. 5, 2011, doi: 10.1002/rob.20407.
- [55] N. Nourani-Vatani, J. Roberts, and M. V. Srinivasan, "Practical visual odometry for car-like vehicles," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2009. doi: 10.1109/ROBOT.2009.5152403.
- [56] R. Gonzalez, F. Rodriguez, J. L. Guzman, C. Pradalier, and R. Siegwart, "Control of off-road mobile robots using visual odometry and slip compensation," *Advanced Robotics*, vol. 27, no. 11, 2013, doi: 10.1080/01691864.2013.791742.
- [57] R. Gonzalez, F. Rodriguez, J. L. Guzman, C. Pradalier, and R. Siegwart, "Combined visual odometry and visual compass for off-road mobile robots localization," *Robotica*, vol. 30, no. 6, 2012, doi: 10.1017/S026357471100110X.
- [58] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif Intell*, vol. 17, no. 1–3, 1981, doi: 10.1016/0004-3702(81)90024-2.

- [59] B. D. Lucas and T. Kanade, "ITERATIVE IMAGE REGISTRATION TECHNIQUE WITH AN APPLICATION TO STEREO VISION.," 1981.
- [60] A. E. Johnson, S. B. Goldberg, Y. Cheng, and L. H. Matthies, "Robust and efficient stereo feature tracking for visual odometry," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2008. doi: 10.1109/ROBOT.2008.4543184.
- [61] M. O. A. Aqel, M. H. Marhaban, M. I. Sariipan, and N. B. Ismail, "Review of visual odometry: types, approaches, challenges, and applications," *SpringerPlus*, vol. 5, no. 1. 2016. doi: 10.1186/s40064-016-3573-7.
- [62] P. Kicman and J. Narkiewicz, "Concept of integrated INS/visual system for autonomous mobile robot operation," in *Marine Navigation and Safety of Sea Transportation: Navigational Problems*, 2013. doi: 10.1201/b14962-6.
- [63] D. Scaramuzza and R. Siegwart, "Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles," *IEEE Transactions on Robotics*, vol. 24, no. 5, 2008, doi: 10.1109/TRO.2008.2004490.
- [64] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014.
- [65] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015. doi: 10.1109/CVPR.2015.7299059.
- [66] J. Donahue *et al.*, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 4, 2017, doi: 10.1109/TPAMI.2016.2599174.
- [67] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.213.
- [68] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016. doi: 10.1007/978-3-319-46484-8_2.
- [69] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. doi: 10.1109/CVPR.2017.502.
- [70] K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00685.
- [71] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018. doi: 10.1007/978-3-030-01267-0_19.
- [72] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local Neural Networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00813.

- [73] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019. doi: 10.1109/ICCV.2019.00630.
- [74] A. Karpathy, A. Joulin, and F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in Neural Information Processing Systems*, 2014.
- [75] B. SravyaPranati, D. Suma, C. ManjuLatha, and S. Putheti, "Large-Scale Video Classification with Convolutional Neural Networks," in *Smart Innovation, Systems and Technologies*, 2021. doi: 10.1007/978-981-15-7062-9_69.
- [76] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends in Neurosciences*, vol. 15, no. 1. 1992. doi: 10.1016/0166-2236(92)90344-8.
- [77] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards Good Practices for Very Deep Two-Stream ConvNets," Jul. 2015, Accessed: Jun. 27, 2023. [Online]. Available: <http://arxiv.org/abs/1507.02159>
- [78] C. Feichtenhofer, A. Pinz, and R. P. Wildes., "Spatiotemporal Residual Networks for Video Action Recognition.," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [79] C. Feichtenhofer, ... A. P.-P. of the I., and undefined 2017, "Spatiotemporal multiplier networks for video action recognition," *openaccess.thecvf.com*, Accessed: Jun. 27, 2023. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2017/html/Feichtenhofer_Spatiotemporal_Multiplier_Networks_CVPR_2017_paper.html
- [80] Y. Wang, M. Long, J. Wang, and P. S. Yu, "Spatiotemporal pyramid network for video action recognition," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. doi: 10.1109/CVPR.2017.226.
- [81] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional neural networks for human action recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 1, 2013, doi: 10.1109/TPAMI.2012.59.
- [82] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks".
- [83] L. Yao *et al.*, "Describing Videos by Exploiting Temporal Structure", Accessed: Jun. 27, 2023. [Online]. Available: <https://www.youtube.com/yt/press/statistics>.
- [84] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. doi: 10.1109/CVPR.2017.634.
- [85] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Multi-fiber Networks for Video Recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018. doi: 10.1007/978-3-030-01246-5_22.
- [86] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 42, no. 8, 2020, doi: 10.1109/TPAMI.2019.2913372.
- [87] A. Diba *et al.*, "Spatio-temporal Channel Correlation Networks for Action Classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018. doi: 10.1007/978-3-030-01225-0_18.

- [88] Z. Qiu, T. Yao, and T. Mei, “Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017. doi: 10.1109/ICCV.2017.590.
- [89] D. Tran, H. Wang, L. Torresani, J. Ray, Y. Lecun, and M. Paluri, “A Closer Look at Spatiotemporal Convolutions for Action Recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00675.
- [90] Y. Zhao, Y. Xiong, and D. Lin, “Trajectory convolution for action recognition,” in *Advances in Neural Information Processing Systems*, 2018.
- [91] “youtube,” <https://www.youtube.com/>.
- [92] Bruce D. Lucas and Takeo Kanade, “Iterative Technique of Image Registration and Its Application to Stereo,” in *Proceedings of the International Joint Conference on Neural Networks*, 1981.
- [93] J. J. Lee and G. Kim, “Robust estimation of camera homography using fuzzy RANSAC,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2007. doi: 10.1007/978-3-540-74472-6_81.
- [94] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *Neural information processing systems*, vol. 1, 2006.
- [95] G. Hinton *et al.*, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” *IEEE Signal Process Mag*, no. November, 2012, doi: 10.1109/MSP.2012.2205597.
- [96] W. Guo, J. Wang, and S. Wang, “Deep Multimodal Representation Learning: A Survey,” *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2916887.
- [97] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, “Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training,” in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020. doi: 10.1609/aaai.v34i07.6795.
- [98] W. Su *et al.*, “VL-BERT: Pre-training of Generic Visual-Linguistic Representations,” Aug. 2019, Accessed: Jul. 02, 2023. [Online]. Available: <https://arxiv.org/abs/1908.08530v4>
- [99] Y. C. Chen *et al.*, “UNITER: UNiversal Image-TExt Representation Learning,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020. doi: 10.1007/978-3-030-58577-8_7.
- [100] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, “VideoBERT: A joint model for video and language representation learning,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019. doi: 10.1109/ICCV.2019.00756.
- [101] B. Korbar, F. Petroni, R. Girdhar, and L. Torresani, “Video Understanding as Machine Translation,” Jun. 2020.
- [102] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in *Advances in Neural Information Processing Systems*, 2019.
- [103] H. Tan and M. Bansal, “LXMert: Learning cross-modality encoder representations from transformers,” in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural*

- Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019. doi: 10.18653/v1/d19-1514.
- [104] S. K. D’Mello and J. Kory, “A review and meta-analysis of multimodal affect detection systems,” *ACM Computing Surveys*, vol. 47, no. 3, 2015. doi: 10.1145/2682899.
- [105] E. Shutova, D. Kiela, and J. Maillard, “Black holes and white rabbits: Metaphor identification with visual features,” in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 2016. doi: 10.18653/v1/n16-1020.
- [106] Z. Z. Lan, L. Bao, S. I. Yu, W. Liu, and A. G. Hauptmann, “Multimedia classification and event detection using double fusion,” *Multimed Tools Appl*, vol. 71, no. 1, 2014, doi: 10.1007/s11042-013-1391-2.
- [107] Z. J. Zha, J. Liu, T. Yang, and Y. Zhang, “Spatiotemporal-textual co-attention network for video question answering,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 15, no. 2s, 2019, doi: 10.1145/3320061.
- [108] K. Xu *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *32nd International Conference on Machine Learning, ICML 2015*, 2015.
- [109] T. Baltrusaitis, C. Ahuja, and L. P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, 2019. doi: 10.1109/TPAMI.2018.2798607.
- [110] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy, “What’s Cookin’? Interpreting Cooking Videos using Text, Speech and Vision,” *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 143–152, Mar. 2015, doi: 10.3115/v1/n15-1015.
- [111] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao, “Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends,” *Foundations and Trends in Computer Graphics and Vision*, vol. 14, no. 3–4, 2022, doi: 10.1561/0600000105.
- [112] A. Dosovitskiy *et al.*, “AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE,” in *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.
- [113] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020. doi: 10.1007/978-3-030-58452-8_13.
- [114] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, “Perceiver: General Perception with Iterative Attention,” in *Proceedings of the 38th International Conference on Machine Learning*, 2021.
- [115] N. Srivastava and R. Salakhutdinov, “Multimodal learning with Deep Boltzmann Machines,” *Journal of Machine Learning Research*, vol. 15, 2014.
- [116] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL HLT 2019 - 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.

- [117] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou, “Training Vision Transformers for Image Retrieval,” Feb. 2021, Accessed: Jul. 17, 2023. [Online]. Available: <https://arxiv.org/abs/2102.05644v1>
- [118] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in Vision: A Survey,” *ACM Comput Surv*, vol. 54, no. 10, 2022, doi: 10.1145/3505244.
- [119] H. Chefer, S. Gur, and L. Wolf, “Transformer Interpretability Beyond Attention Visualization,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021. doi: 10.1109/CVPR46437.2021.00084.
- [120] S. Abnar and W. Zuidema, “Quantifying Attention Flow in Transformers,” May 2020.
- [121] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, “Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned,” May 2019.
- [122] H. Touvron, F. Massa, M. Cord, and A. Sablayrolles, “Training data-efficient image transformers & distillation through attention arXiv : 2012 . 12877v2 [cs . CV] 15 Jan 2021,” *ArXiv*, 2021.
- [123] Z. Gan *et al.*, “Playing Lottery Tickets with Vision and Language,” *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, vol. 36, pp. 652–660, Apr. 2021, doi: 10.1609/aaai.v36i1.19945.
- [124] R. Child, S. Gray, A. Radford, and I. Sutskever, “Generating Long Sequences with Sparse Transformers,” Apr. 2019, Accessed: Jul. 02, 2023. [Online]. Available: <https://arxiv.org/abs/1904.10509v1>
- [125] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018. doi: 10.18653/v1/p18-1238.
- [126] T. Y. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014. doi: 10.1007/978-3-319-10602-1_48.
- [127] A. Agrawal *et al.*, “VQA: Visual Question Answering: www.visualqa.org,” *Int J Comput Vis*, vol. 123, no. 1, 2017, doi: 10.1007/s11263-016-0966-6.
- [128] X. Zhan *et al.*, “Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal Pretraining,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.01157.
- [129] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, “Pano-AVQA: Grounded Audio-Visual Question Answering on 360° Videos,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00204.
- [130] P. Morgado, Y. Li, and N. Vasconcelos, “Learning Representations from Audio-Visual Spatial Alignment,” *Adv Neural Inf Process Syst*, vol. 2020-December, Nov. 2020, Accessed: Jul. 02, 2023. [Online]. Available: <https://arxiv.org/abs/2011.01819v1>

- [131] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, “AI Choreographer: Music Conditioned 3D Dance Generation with AIST++,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.01315.
- [132] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, “Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00213.
- [133] R. Sawhney, M. Goyal, P. Goel, P. Mathur, and R. R. Shah, “Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines,” in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2021. doi: 10.18653/v1/2021.acl-long.526.
- [134] J. Zhang, M. Zheng, M. Boyd, and E. Ohn-Bar, “X-World: Accessibility, Vision, and Autonomy Meet,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00962.
- [135] D. Zhang, M. Zhang, H. Zhang, L. Yang, and H. Lin, “MultiMET: A multimodal dataset for metaphor understanding,” in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2021. doi: 10.18653/v1/2021.acl-long.249.
- [136] D. Kiela *et al.*, “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes,” *Adv Neural Inf Process Syst*, vol. 2020-December, May 2020, Accessed: Jul. 02, 2023. [Online]. Available: <https://arxiv.org/abs/2005.04790v3>
- [137] A. Miech, Di. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, “HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019. doi: 10.1109/ICCV.2019.00272.
- [138] L. Zhou, C. Xu, and J. J. Corso, “Towards automatic learning of procedures from web instructional videos,” in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018. doi: 10.1609/aaai.v32i1.12342.
- [139] Z. Yu *et al.*, “ActivityNet-QA: A dataset for understanding complex web videos via question answering,” in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 2019. doi: 10.1609/aaai.v33i01.33019127.
- [140] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, “TGIF-QA: Toward spatio-temporal reasoning in visual question answering,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. doi: 10.1109/CVPR.2017.149.
- [141] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, “Just Ask: Learning to Answer Questions from Millions of Narrated Videos,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00171.
- [142] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, “MovieQA: Understanding stories in movies through question-answering,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.501.

- [143] J. Lei, L. Yu, M. Bansal, and T. L. Berg, “TVQA: Localized, compositional video question answering,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2018. doi: 10.18653/v1/d18-1167.
- [144] Z. Fang, T. Gokhale, P. Banerjee, C. Baral, and Y. Yang, “Video2Commonsense: Generating commonsense descriptions to enrich video captioning,” in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020. doi: 10.18653/v1/2020.emnlp-main.61.
- [145] J. Li, L. Niu, and L. Zhang, “From Representation to Reasoning: Towards both Evidence and Commonsense Reasoning for Video Question-Answering,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.02059.
- [146] X. Song, Y. Shi, X. Chen, and Y. Han, “Explore multi-step reasoning in video question answering,” in *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, 2018. doi: 10.1145/3240508.3240563.
- [147] A. Zadeh, M. Chan, P. P. Liang, E. Tong, and L. P. Morency, “Social-IQ: A question answering benchmark for artificial social intelligence,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.00901.
- [148] T. Maharaj, N. Ballas, A. Rohrbach, A. Courville, and C. Pal, “A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. doi: 10.1109/CVPR.2017.778.
- [149] K. Yi *et al.*, “CLEVRER: COLLISION EVENTS FOR VIDEO REPRESENTATION AND REASONING,” in *8th International Conference on Learning Representations, ICLR 2020*, 2020.
- [150] N. Garcia, M. Otani, C. Chu, and Y. Nakashima, “KnowIT VQA: Answering knowledge-based questions about videos,” in *AAAI 2020 - 34th AAI Conference on Artificial Intelligence*, 2020. doi: 10.1609/aaai.v34i07.6713.
- [151] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.
- [152] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” *Proceedings of the workshop on text summarization branches out (WAS 2004)*, no. 1, 2004.
- [153] A. Lavie and A. Agarwal, “METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2007.
- [154] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *Journal of Artificial Intelligence Research*, vol. 47, 2013, doi: 10.1613/jair.3994.
- [155] Z. Wu and M. Palmer, “Verb semantics and lexical selection,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1994. doi: 10.3115/981732.981751.
- [156] G. A. Miller, “WordNet: A Lexical Database for English,” *Commun ACM*, vol. 38, no. 11, 1995, doi: 10.1145/219717.219748.

- [157] M. Malinowski and M. Fritz, “A multi-world approach to question answering about real-world scenes based on uncertain input,” in *Advances in Neural Information Processing Systems*, 2014.
- [158] X. Chen *et al.*, “Microsoft COCO Captions: Data Collection and Evaluation Server,” Apr. 2015, Accessed: Jul. 05, 2023. [Online]. Available: <https://arxiv.org/abs/1504.00325v2>
- [159] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “ViViT: A Video Vision Transformer,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00676.
- [160] K. H. Zeng, T. H. Chen, C. Y. Chuang, Y. H. Liao, J. C. Niebles, and M. Sun, “Leveraging video descriptions to learn video question answering,” in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017. doi: 10.1609/aaai.v31i1.11238.
- [161] Z. Zhao, J. Lin, X. Jiang, D. Cai, X. He, and Y. Zhuang, “Video question answering via hierarchical dual-level attention network learning,” in *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, 2017. doi: 10.1145/3123266.3123364.
- [162] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang, “Video question answering via hierarchical spatio-temporal attention networks,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2017. doi: 10.24963/ijcai.2017/492.
- [163] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *30th International Conference on Machine Learning, ICML 2013*, 2013.
- [164] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, “End-to-end memory networks,” in *Advances in Neural Information Processing Systems*, 2015.
- [165] S. Na, S. Lee, J. Kim, and G. Kim, “A Read-Write Memory Network for Movie Story Understanding,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 677–685, Sep. 2017, doi: 10.1109/ICCV.2017.80.
- [166] K. M. Kim, M. O. Heo, S. H. Choi, and B. T. Zhang, “Deepstory: Video story QA by deep embedded memory networks,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2017. doi: 10.24963/ijcai.2017/280.
- [167] J. Gao, R. Ge, K. Chen, and R. Nevatia, “Motion-Appearance Co-memory Networks for Video Question Answering,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00688.
- [168] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, “Heterogeneous memory enhanced multimodal attention model for video question answering,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.00210.
- [169] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.
- [170] P. Jiang and Y. Han, “Reasoning with heterogeneous graph alignment for video question answering,” in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020. doi: 10.1609/aaai.v34i07.6767.
- [171] J. Park, J. Lee, and K. Sohn, “Bridge to Answer: Structure-aware Graph Interaction Network for Video Question Answering,” 2021. doi: 10.1109/cvpr46437.2021.01527.

- [172] J. Wang, B. K. Bao, and C. Xu, “DualVGR: A Dual-Visual Graph Reasoning Unit for Video Question Answering,” *IEEE Trans Multimedia*, vol. 24, 2022, doi: 10.1109/TMM.2021.3097171.
- [173] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, “Location-aware graph convolutional networks for video question answering,” in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020. doi: 10.1609/aaai.v34i07.6737.
- [174] X. Li *et al.*, “Beyond RNNs: Positional self-attention with co-attention for video question answering,” in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 2019. doi: 10.1609/aaai.v33i01.33018658.
- [175] J. Xiao, A. Yao, Z. Liu, Y. Li, W. Ji, and T. S. Chua, “Video as Conditional Graph Hierarchy for Multi-Granular Question Answering,” in *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, 2022. doi: 10.1609/aaai.v36i3.20184.
- [176] L. Peng, S. Yang, Y. Bin, and G. Wang, “Progressive Graph Attention Network for Video Question Answering,” in *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, 2021. doi: 10.1145/3474085.3475193.
- [177] F. Liu, J. Liu, W. Wang, and H. Lu, “HAIR: Hierarchical Visual-Semantic Relational Reasoning for Video Question Answering,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00172.
- [178] F. L. Chen *et al.*, “VLP: A Survey on Vision-Language Pre-training,” *Machine Intelligence Research*, vol. 20, no. 1, pp. 38–56, Feb. 2022, doi: 10.1007/s11633-022-1369-5.
- [179] L. Li, Y. C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, “HERO: Hierarchical encoder for video+language omni-representation pre-training,” in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020. doi: 10.18653/v1/2020.emnlp-main.161.
- [180] J. Lei *et al.*, “Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021. doi: 10.1109/CVPR46437.2021.00725.
- [181] R. Krishna *et al.*, “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations,” *Int J Comput Vis*, vol. 123, no. 1, 2017, doi: 10.1007/s11263-016-0981-7.
- [182] R. Zellers *et al.*, “MERLOT: Multimodal Neural Script Knowledge Models,” in *Advances in Neural Information Processing Systems*, 2021.
- [183] R. Zellers *et al.*, “MERLOT RESERVE: Neural Script Knowledge through Vision and Language and Sound,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01589.
- [184] T.-J. Fu *et al.*, “VIOLET : End-to-End Video-Language Transformers with Masked Visual-token Modeling,” Nov. 2021, Accessed: Jul. 05, 2023. [Online]. Available: <https://arxiv.org/abs/2111.12681v2>

- [185] Z. Liu *et al.*, “Video Swin Transformer,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.00320.
- [186] A. Ramesh *et al.*, “Zero-Shot Text-to-Image Generation,” Feb. 2021, Accessed: Jul. 05, 2023. [Online]. Available: <https://arxiv.org/abs/2102.12092v2>
- [187] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, “Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00175.
- [188] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, and H. Takemura, “BERT representations for video question answering,” in *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, 2020. doi: 10.1109/WACV45572.2020.9093596.
- [189] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 6, 2017, doi: 10.1109/TPAMI.2016.2577031.
- [190] D. Engin, F. Schnitzler, N. Q. K. Duong, and Y. Avrithis, “On the hidden treasure of dialog in video question answering,” 2022. doi: 10.1109/iccv48922.2021.00207.
- [191] M. Lewis *et al.*, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.703.
- [192] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks,” in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019. doi: 10.18653/v1/d19-1410.
- [193] W. Yu *et al.*, “Learning from Inside: Self-driven Siamese Sampling and Reasoning for Video Question Answering,” in *Advances in Neural Information Processing Systems*, 2021.
- [194] A. Miech, J. B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, “End-to-End Learning of Visual Representations from Uncurated Instructional Videos,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9876–9886, Dec. 2019, doi: 10.1109/CVPR42600.2020.00990.
- [195] D. Hendrycks and K. Gimpel, “Gaussian Error Linear Units (GELUs),” Jun. 2016, Accessed: Jul. 02, 2023. [Online]. Available: <https://arxiv.org/abs/1606.08415v5>
- [196] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Dec. 2022, Accessed: Jun. 16, 2023. [Online]. Available: <https://arxiv.org/abs/2212.04356v1>
- [197] Y. Wu *et al.*, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” Sep. 2016, Accessed: Jul. 03, 2023. [Online]. Available: <https://arxiv.org/abs/1609.08144v2>
- [198] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” Oct. 2019, Accessed: Jun. 30, 2023. [Online]. Available: <https://arxiv.org/abs/1910.01108v4>
- [199] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” *Psychol Rev*, vol. 65, no. 6, 1958, doi: 10.1037/h0042519.

- [200] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *Proceedings - IEEE Symposium on Security and Privacy*, 2017. doi: 10.1109/SP.2017.49.
- [201] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
- [202] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.282.
- [203] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," in *IJCAI International Joint Conference on Artificial Intelligence*, 2018. doi: 10.24963/ijcai.2018/585.
- [204] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings - 2016 IEEE European Symposium on Security and Privacy, EURO S and P 2016*, 2016. doi: 10.1109/EuroSP.2016.36.
- [205] N. Papernot, P. McDaniel, A. Swami, and R. Harang, "Crafting adversarial input sequences for recurrent neural networks," in *Proceedings - IEEE Military Communications Conference MILCOM*, 2016. doi: 10.1109/MILCOM.2016.7795300.
- [206] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial examples for malware detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017. doi: 10.1007/978-3-319-66399-9_4.
- [207] Y. Belinkov and Y. Bisk, "Synthetic and natural noise both break neural machine translation," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.
- [208] J. Gao, J. Lanchantin, M. Lou Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*, 2018. doi: 10.1109/SPW.2018.00016.
- [209] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "TextBugger: Generating Adversarial Text Against Real-world Applications," 2019. doi: 10.14722/ndss.2019.23138.
- [210] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2018. doi: 10.18653/v1/n18-1170.
- [211] M. T. Ribeiro, S. Singh, and C. Guestrin, "Summery 2," *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, 2018.
- [212] Z. Zhao, D. Dua, and S. Singh, "Generating natural adversarial examples," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.
- [213] H. Gao, H. Zhang, X. Yang, W. Li, F. Gao, and Q. Wen, "Generating natural adversarial examples with universal perturbations for text classification," *Neurocomputing*, vol. 471, 2022, doi: 10.1016/j.neucom.2021.10.089.

- [214] I. J. Goodfellow *et al.*, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, 2014. doi: 10.1007/978-3-658-40442-0_9.
- [215] H. Chen, H. Zhang, P. Y. Chen, J. Yi, and C. J. Hsieh, “Attacking visual language grounding with adversarial examples: A case study on neural image captioning,” in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018. doi: 10.18653/v1/p18-1241.
- [216] C. Song and V. Shmatikov, “Fooling OCR Systems with Adversarial Text Images,” Feb. 2018, Accessed: Jul. 03, 2023. [Online]. Available: <https://arxiv.org/abs/1802.05385v1>
- [217] X. Xu, X. Chen, C. Liu, A. Rohrbach, T. Darrell, and D. Song, “Fooling Vision and Language Models Despite Localization and Attention Mechanism,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00520.
- [218] A. Fan, M. Lewis, and Y. Dauphin, “Hierarchical neural story generation,” in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018. doi: 10.18653/v1/p18-1082.
- [219] N. Ansari and R. Sharma, “Identifying Semantically Duplicate Questions Using Data Science Approach: A Quora Case Study,” vol. 11, Apr. 2020, Accessed: Jul. 03, 2023. [Online]. Available: <https://arxiv.org/abs/2004.11694v1>
- [220] W. B. Dolan and C. Brockett, “Automatically Constructing a Corpus of Sentential Paraphrases,” *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.
- [221] Y. Zhang, J. Baldridge, and L. He, “PAWS: Paraphrase Adversaries from Word Scrambling,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 1298–1308, Apr. 2019, Accessed: Jul. 03, 2023. [Online]. Available: <https://arxiv.org/abs/1904.01130v1>
- [222] T. M. Le, V. Le, S. Venkatesh, and T. Tran, “Hierarchical Conditional Relation Networks for Multimodal Video Question Answering,” *Int J Comput Vis*, vol. 129, no. 11, 2021, doi: 10.1007/s11263-021-01514-3.
- [223] P. H. Seo, A. Nagrani, and C. Schmid, “Look before you speak: Visually contextualized utterances,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021. doi: 10.1109/CVPR46437.2021.01660.
- [224] A. Kojima, T. Tamura, and K. Fukunaga, “Natural language description of human activities from video images based on concept hierarchy of actions,” *Int J Comput Vis*, vol. 50, no. 2, 2002, doi: 10.1023/A:1020346032608.
- [225] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, “Translating video content to natural language descriptions,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013. doi: 10.1109/ICCV.2013.61.
- [226] J. Zhang and Y. Peng, “Object-aware aggregation with bidirectional temporal graph for video captioning,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.00852.

- [227] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to Sequence -- Video to Text,” in *Proceedings of the IEEE international conference on computer vision.*, May 2015.
- [228] X. Zhang, C. Liu, and F. Chang, “Guidance Module Network for Video Captioning,” in *Chinese Control Conference, CCC*, 2021. doi: 10.23919/CCC52363.2021.9550288.
- [229] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-ResNet and the impact of residual connections on learning,” in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017. doi: 10.1609/aaai.v31i1.11231.
- [230] H. Xu, B. Li, V. Ramanishka, L. Sigal, and K. Saenko, “Joint event detection and description in continuous video streams,” in *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, 2019. doi: 10.1109/WACV.2019.00048.
- [231] M. Hemalatha and C. C. Sekhar, “Domain-specific semantics guided approach to video captioning,” in *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, 2020. doi: 10.1109/WACV45572.2020.9093344.
- [232] K. Hara, H. Kataoka, and Y. Satoh, “Learning spatio-Temporal features with 3D residual networks for action recognition,” in *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, 2017. doi: 10.1109/ICCVW.2017.373.
- [233] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [234] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014. doi: 10.3115/v1/d14-1179.
- [235] Z. Guo, L. Gao, J. Song, X. Xu, J. Shao, and H. T. Shen, “Attention-based LSTM with semantic consistency for videos captioning,” in *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, 2016. doi: 10.1145/2964284.2967242.
- [236] Z. Zhang *et al.*, “Object relational graph with teacher-recommended learning for video captioning,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/CVPR42600.2020.01329.
- [237] J. Hessel, B. Pang, Z. Zhu, and R. Soricut, “A case study on combining ASR and visual features for generating instructional video captions,” in *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, 2019. doi: 10.18653/v1/k19-1039.
- [238] B. Shi *et al.*, “Dense procedure captioning in narrated instructional videos,” in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/p19-1641.
- [239] H. Xu *et al.*, “VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding,” Sep. 2021.
- [240] H. Xu *et al.*, “mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video,” in *ICML2023*, Feb. 2023.
- [241] Y. Fang *et al.*, “EVA: Exploring the Limits of Masked Visual Representation Learning at Scale,” Nov. 2022.

- [242] K. Lin *et al.*, “SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning,” Nov. 2021.
- [243] H. Luo *et al.*, “UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation,” Feb. 2020, Accessed: Jun. 30, 2023. [Online]. Available: <https://arxiv.org/abs/2002.06353v3>
- [244] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, “End-to-end Generative Pretraining for Multimodal Video Captioning,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01743.
- [245] H. Xu *et al.*, “mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video,” Feb. 2023.
- [246] S. Chen *et al.*, “VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset,” May 2023.
- [247] S. Yan *et al.*, “VideoCoCa: Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners,” Dec. 2022.
- [248] B. Yang, T. Zhang, and Y. Zou, “CLIP Meets Video Captioning: Concept-Aware Representation Learning Does Matter,” Nov. 2021.
- [249] X. Wang *et al.*, “Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey,” Feb. 2023.
- [250] J. Wang *et al.*, “GIT: A Generative Image-to-text Transformer for Vision and Language,” May 2022.
- [251] S. Banerjee and A. Lavie, “METEOR: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Proceedings of the Workshop ACL 2005*, 2005.
- [252] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.
- [253] Z. Zhang *et al.*, “Open-book Video Captioning with Retrieve-Copy-Generate Network,” Mar. 2021.

Résumé en français

Titre : Analyse multimodale par apprentissage profond pour la production audiovisuelle

Mots clés : Indexation d'archives, Apprentissage multimodal, Traitement du langage naturel, Vision par ordinateur

Résumé : Dans le contexte en constante évolution du contenu audiovisuel, la nécessité cruciale d'automatiser l'indexation et l'organisation des archives s'est imposée comme un objectif primordial. En réponse, cette recherche explore l'utilisation de techniques d'apprentissage profond pour automatiser l'extraction de métadonnées diverses dans les archives, améliorant ainsi leur accessibilité et leur réutilisation.

La première contribution de cette recherche tourne autour de la classification des types de mouvements de caméra. Il s'agit d'un aspect crucial de l'indexation du contenu, car il permet une catégorisation efficace et une récupération du contenu vidéo en fonction de la dynamique visuelle qu'il présente. L'approche novatrice proposée utilise des réseaux neuronaux convolutionnels 3D avec des blocs résiduels. Une approche semi-automatique pour la construction d'un ensemble de données fiable sur les mouvements de caméra à partir de vidéos disponibles au public est également présentée, réduisant au minimum le besoin d'intervention manuelle. De plus, la création d'un ensemble de données d'évaluation exigeant, comprenant des vidéos de la vie réelle tournées avec des caméras professionnelles à différentes résolutions, met en évidence la robustesse et la capacité de généralisation de la technique proposée, atteignant un taux de précision moyen de 94 %.

La deuxième contribution se concentre sur la tâche de *Video Question Answering*. Dans ce contexte, nous explorons l'efficacité des *transformers* basés sur l'attention pour faciliter l'apprentissage multimodal ancré. Le défi ici réside dans le comblement de l'écart entre les modalités visuelles et textuelles et dans la réduction de la complexité quadratique des modèles de *transformers*. Pour résoudre ces problèmes, un nouveau cadre est introduit, qui intègre un *transformers* léger et un module de cross-modalité. Ce module utilise une corrélation croisée pour permettre un apprentissage réciproque entre les caractéristiques visuelles conditionnées par le texte et les caractéristiques textuelles conditionnées par la vidéo. De plus, un scénario de test adversarial avec des questions reformulées met en évidence la robustesse du modèle et son applicabilité dans le monde réel. Les résultats expérimentaux sur des ensembles de données de référence, tels que MSVD-QA et MSRVT-QA, valident la méthodologie proposée, avec une précision moyenne de 45 % et 42 % respectivement, ce qui représente des améliorations notables par rapport aux approches existantes.

La troisième contribution de cette recherche aborde le problème de *video captioning*, un aspect critique de l'indexation du contenu. Le travail introduit intègre un module de *modality attention* qui capture les relations complexes entre les données visuelles et textuelles à l'aide d'une corrélation croisée. De plus, l'intégration de l'attention temporelle améliore la capacité du modèle à produire des légendes significatives en tenant compte de la dynamique temporelle du contenu vidéo. Notre travail intègre également une tâche auxiliaire utilisant une fonction de perte contrastive, ce qui favorise la généralisation du modèle et une compréhension plus approfondie des relations intermodales et des sémantiques sous-jacentes. L'utilisation d'une architecture de

transformer pour l'encodage et le décodage améliore considérablement la capacité du modèle à capturer les interdépendances entre les données textuelles et vidéo. La recherche valide la méthodologie proposée par une évaluation rigoureuse sur la référence MSRVT, atteignant des scores BLEU4, ROUGE et METEOR de 0,4408, 0,6291 et 0,3082 respectivement. En comparaison avec les méthodes de l'état de l'art, cette approche surpasse de manière constante, avec des gains de performance allant de 1,21 % à 1,52 % pour les trois métriques considérées.

En conclusion, ce manuscrit offre une exploration holistique des techniques basées sur l'apprentissage profond pour automatiser l'indexation du contenu télévisuel, en abordant la nature laborieuse et chronophage de l'indexation manuelle. Les contributions englobent la classification des types de mouvements de caméra, la *video question answering* et la *video captioning*, faisant avancer collectivement l'état de l'art et fournissant des informations précieuses pour les chercheurs dans le domaine. Ces découvertes ont non seulement des applications pratiques pour la recherche et l'indexation de contenu, mais contribuent également à l'avancement plus large des méthodologies d'apprentissage profond dans le contexte multimodal.

Abstract

Title : Multimodal analysis using deep learning techniques for audiovisual production

Keywords : Archive indexing, Multimodal learning, Natural Language Processing, Computer vision, Video captioning, Video question answering.

Abstract : Within the dynamic landscape of television content, the critical need to automate the indexing and organization of archives has emerged as a paramount objective. In response, this research explores the use of deep learning techniques to automate the extraction of diverse metadata from television archives, improving their accessibility and reuse.

The first contribution of this research revolves around the classification of camera motion types. This is a crucial aspect of content indexing as it allows for efficient categorization and retrieval of video content based on the visual dynamics it exhibits. The novel approach proposed employs 3D convolutional neural networks with residual blocks, a technique inspired by action recognition methods. A semi-automatic approach for constructing a reliable camera motion dataset from publicly available videos is also presented, minimizing the need for manual intervention. Additionally, the creation of a challenging evaluation dataset, comprising real-life videos shot with professional cameras at varying resolutions, underlines the robustness and generalization power of the proposed technique, achieving an average accuracy rate of 94%.

The second contribution centers on the demanding task of Video Question Answering. In this context, we explore the effectiveness of attention-based transformers for facilitating grounded multimodal learning. The challenge here lies in bridging the gap between the visual and textual modalities and mitigating the quadratic complexity of transformer models. To address these issues, a novel framework is introduced, which incorporates a lightweight transformer and a cross-modality module. This module leverages cross-correlation to enable reciprocal learning between text-conditioned visual features and video-conditioned textual features. Furthermore, an adversarial testing scenario with rephrased questions highlights the model's robustness and real-world applicability. Experimental results on benchmark datasets, such as MSVD-QA and MSRVTT-QA, validate the proposed methodology, with an average accuracy of 45% and 42%, respectively, which represents notable improvements over existing approaches.

The third contribution of this research addresses the multimodal video captioning problem, a critical aspect of content indexing. The introduced framework incorporates a modality-attention module that captures the intricate relationships between visual and textual data using cross-correlation. Moreover, the integration of temporal attention enhances the model's ability to produce meaningful captions, considering the temporal dynamics of video content. Our work also incorporates an auxiliary task employing a contrastive loss function, which promotes model generalization and a deeper understanding of inter-modal relationships and underlying semantics. The utilization of a transformer architecture for encoding and decoding significantly enhances the model's capacity to capture interdependencies between text and video data. The research validates the proposed methodology through rigorous evaluation on the MSRVTT benchmark,

achieving BLEU4, ROUGE, and METEOR scores of 0.4408, 0.6291 and 0.3082, respectively. In comparison to state-of-the-art methods, this approach consistently outperforms, with performance gains ranging from 1.21% to 1.52% across the three metrics considered.

In conclusion, this manuscript offers a holistic exploration of deep learning-based techniques to automate television content indexing, addressing the labor-intensive and time-consuming nature of manual indexing. The contributions encompass camera motion type classification, VideoQA, and multimodal video captioning, collectively advancing the state of the art and providing valuable insights for researchers in the field. These findings not only have practical applications for content retrieval and indexing but also contribute to the broader advancement of deep learning methodologies in the multimodal context.

