A Thesis for the Degree of Ph.D. in Engineering

Performance Efficiency in Multimodal Representation Learning

January 2024

Graduate School of Science and Technology
Keio University

Yanan Wang

Abstract

Large-scale foundation models have shown promising success in processing multimodal information (*e.g.*, vision, text, audio, etc.) for the purpose of achieving artificial general intelligence. However, conventional systems are typically built on the assumption that all modalities exist, so the lack of modalities will lead to poor multimodal fusion, which results in lower inference performance and potentially limits the capability in practical applications. In addition, multimodal systems require tremendous parameters, making inference computationally inefficient. Therefore, studying how to improve the performance efficiency of multimodal representation learning is crucial to achieving realistic multimodal systems.

This thesis involves various endeavors aimed at improving the performance efficiency of multimodal representation learning. These efforts not only facilitate high-performance multimodal fusion but also enhance the efficiency of multimodal inference. This thesis consists of two parts. In part 1, we focus on performing high-performance multimodal fusion. We propose a multiple attention fusion network (MAFN), which includes a multimodal domain adaptation module designed to obtain expressive multimodal representations capturing dynamic interactions across modalities. We also propose a VAE-based adversarial multimodal domain transfer (VAE-AMDT) method to further improve effective multimodal fusion by reducing the distribution difference between diverse modal representations. We further propose a new visual question answering (VQA) method called VQA-GNN that performs bidirectional fusion between unstructured and structured multimodal knowledge to obtain unified knowledge representations. In part 2, we focus on achieving effective multimodal transfer learning aiming for efficient multimodal systems—unimodal systems can achieve competitive performance with multimodal systems even when provided with a specific unimodal input. We propose a language knowledge injectable deep neural network (LDNN) to achieve a vision modal model enhanced with implicit language knowledge. Then, we develop VideoAdviser, a video knowledge distillation method to transfer multimodal knowledge from a multimodal fundamental model to a specific modal fundamental model via optimizing a step-distillation objective loss.

Acknowledgements

Four years have passed since I started my Ph.D. program in the fall of 2019. I am so excited and proud that I can conclude my journey at Keio University. Looking back on my journey as an international student, I went to Japan in the spring of 2009 after graduating from high school in China. After spending two years in a Japanese language school (Tokyo Sanritsu Academy), I entered Aoyama Gakuin University for my undergraduate program. Under the guidance of my adviser Yoshihito Tobe, it influenced not only the way I conduct research but also the way I think about my career. Without strong encouragement from my advisor Yoshihito Tobe, I could never have encountered my current advisor Satoshi Kurihara. It is my great honor to have such an opportunity to study more things from my advisor Satoshi Kurihara. After obtaining a master's degree at the University of Electro-Communications, I joined my current company KDDI Corporation. However, I have never forgotten the splendid research experiments in the flexible Kurihara laboratory, which also affected my career in KDDI Research. That is also why I decided to start my Ph.D. program even though I know that is supposed to be very hard for me. Because I need to work for my company's KPI while making progress in my Ph.D. program. I am deeply grateful to many people surrounding me who supported me through my Ph.D. experience. Without them, I would never accomplished my Ph.D. degree.

First, I would like to thank my advisor, Satoshi Kurihara. I learned lots of things from Prof. Kurihara: How to think about problems deeply, how to collaborate with the company to enhance the research result, and how to enjoy research life. Prof. Kurihara always respects my research interests and supports my research. I always have a great time to share my research progress without any stress. Without Prof. Kurihara's great support, I would never completed my Ph.D..

I would also like to thank my advisor Yoshihito Tobe, who was my adviser during my college days. Prof. Tobe let me explore my research interest in machine learning during the last year of my bachelor's program, which helped me find my current research interest in artificial intelligence. Prof. Tobe also encouraged me to pursue a graduate degree to upgrade my future career.

I am thankful to my colleagues in KDDI Research. Especially, my former manager Jianming Wu, and current manager Shinya Wada, who gave me great support in adjusting my work plan. Without their support, I will spend more time completing my thesis.

I thank Prof. Jure Leskovec who gave great insight for improving my research during my visiting scholar at Stanford University. I also thank Ph.D. students Michihiro Yasunaga and Hongyu Ren who gave me tremendous support during the collaborations.

I am also thankful to the committee members, Prof. Shu Yamada, Prof. Yoshihisa Shinozawa, and Prof. Komei Sugiura of the School of Science for Open and Environmental Systems at Keio Univerisity. For their invaluable insights and abundant advice, which greatly contributed to the improvement of my thesis.

I express my sincere gratitude to the Hasegawa International Scholarship Foundation and the Nitori International Scholarship Foundation. Their generous financial support played an invaluable role in facilitating my bachelor's and master's programs respectively.

Finally, I extend my heartfelt gratitude to my family. My wife, Juan Kang, made the selfless choice of prioritizing our family and looking after our children, enabling me to dedicate myself to my career and research. Her unwavering dedication has been instrumental in my pursuits. I am also deeply appreciative of my two daughters, who consistently infuse my research journey with boundless joy. My parents have always encouraged me to pursue my Ph.D.. They believed in my ability more than anyone else. I thank them for their endless love and support.

Contents

Abstract **iv**

Acknowledgements **vi**

1 Introduction **1**

1.1 Motivation 1

1.2 Thesis Outline 3

1.3 Contributions 4

1.4 Bibliographic Remarks 5

Part 1: Improving high-performance multimodal fusion **7**

2 Multi-Attention Fusion Network for Video-based Emotion Recognition **8**

2.1 Introduction 8

2.2 Related Works 10

2.3 Multi-Attention Fusion Network (MAFN) 12

2.3.1 Intra-modality attention mechanism 13

2.3.2 Inter-modality attention mechanism 13

2.3.3 Multimodal domain adaptation module 15

2.4 Experiment 16

2.4.1 Data preprocessing 16

2.4.2 Training setting 17

2.4.3 Result 17

2.4.4 Discussion 17

2.5 Conclusions 19

3 VAE-Based Adversarial Multimodal Domain Transfer for Video-Level Sentiment Analysis 20

3.1 Introduction 20

3.2 Related work 23

3.3 Problem Statement 24

3.4 Modality data preprocessing 24

3.5 Methodology 25

3.5.1 Multi-attention module 25

3.5.2 VAE-AMDT 27

3.5.3 Learning 28

3.6 Experiment 29

3.6.1 Dataset 29

3.6.2 Metric 30

3.6.3 Full model hyperparameters 30

3.6.4 Performance 30

3.6.5 Effect of VAE-AMDT 32

3.6.6 Ablation study 34

3.6.7 Case study 35

3.7 Conclusion 36

4 VQA-GNN: Reasoning with Multimodal Knowledge via Graph Neural Networks for Visual Question Answering 37

4.1 Introduction 37

4.2 Problem Setup 40

4.3 Related Work 40

4.3.1 Multimodal transformer 40

4.3.2 Structured knowledge-based VQA 41

4.4 Methodology 42

4.4.1 Multimodal semantic graph 42

4.4.2 Multimodal GNN-based bidirectional fusion 45

4.4.3 Inference and Learning 46

4.5 Experiments 47

4.5.1 Experiment Setup 47

4.5.2 Performance 48

4.5.3 Comparison with baselines pretrained only on VCR dataset . 53

4.5.4 Comparison results on different question types 55
4.6 Interpretability 55
4.7 Conclusion 57

Part 2: Achieving effective multimodal transfer learning 58

5 LDNN: Linguistic Knowledge Injectable Deep Neural Network for Group Cohesiveness Understanding 59
5.1 Introduction 60
5.2 Related works 62
5.3 Method 63
5.3.1 Visual Encoder and Language Encoder 64
5.3.2 Domain adaptation and linguistic knowledge transition 66
5.4 Experiment 68
5.4.1 Dataset, settings and metric 68
5.4.2 Results 70
5.4.3 Analysis 71
5.5 Conclusion 74

6 Implicit Knowledge Injectable Cross Attention Audiovisual Model for Group Emotion Recognition 75
6.1 Introduction 75
6.2 Related works 78
6.3 Method 79
6.3.1 Audiovisual subnetwork 79
6.3.2 K-injection subnetwork 81
6.3.3 Training and Inference 83
6.4 Experiment 83
6.4.1 Task, dataset 83
6.4.2 Baseline 84
6.4.3 Experimental settings 85
6.4.4 Ensemble strategy 85
6.4.5 Results on the validation set 85
6.4.6 Ensemble results on the testing set 87
6.5 Conclusion 89

7 VideoAdviser: Video Knowledge Distillation for Multimodal Transfer Learning 90
7.1 Introduction . . . 91
7.2 Related work . . . 94
7.3 Problem setting . . . 95
7.4 Methodology . . . 96
7.4.1 The CLIP-based teacher model . . . 96
7.4.2 The RoBERTa-based student model . . . 99
7.4.3 Training objectives . . . 99
7.5 Experiment . . . 100
7.5.1 Dataset . . . 100
7.5.2 Evaluation metric . . . 102
7.5.3 Training setting . . . 102
7.5.4 Performance . . . 102
7.5.5 Efficiency . . . 104
7.5.6 Analysis . . . 105
7.5.7 Significance Testing . . . 110
7.6 Conclusion . . . 112

8 Conclusion 113

Bibliography 115

List of Tables

2.1 Comparison results of MAFN built using various visual encoders. Here, we use the modified validation and test datasets to evaluate the accuracy of the emotion recognition task. 19

2.2 Recognition accuracy of each MAFN w/o multimodal domain adaptation module on the modified dataset. The performance is reduced compared to the results in Tab. 2.1. 19

3.1 The size of datasets. 29

3.2 Full hyperparameters for our model. "dim" indicates the number of dimensions. "Self", "Cross" and "Triple" indicate self-attention, cross-attention, and triple-attention components respectively. 31

3.3 Comparison of *VAE-AMDT* and state-of-the-art results in both MOSI and MOSEI. *VAE-AMDT* outperforms state-of-the-art Self-MM (*MAE/F1*) by over ***0.16/3.6*** point (MOSI) and ***0.05/2.9*** point (MOSEI). Here, the lower the MAE, the better the performance. (*) indicates that the results are referenced from the Self-MM paper. (+) indicates that Self-MM is trained by using the same preprocessed data in our method; (CV) indicates the result of the 10-fold cross validation. . . . 32

3.4 Comparison results of the model trained w/o and w/ VAE-AMDT. The model trained with VAE-AMDT further improves F1 score of (w/o VAE-AMDT) by 3.6% (MOSI) and 1.7% (MOSEI). 32

3.5 MMD results show that not only can the model (w/ VAE-AMDT) balance the distance between any two modalities, but the average result is lower than the model (w/ AMDT). 33

3.6 Ablation study of the multi-attention module on MOSEI dataset. All models are trained w/o VAE-ADMT, (self, cross, triple)-attention shows the lowest MAE score compared to others. 35

3.7 Effect of modality. The test results show that adding modality improves performance. 35

3.8 Case study on MOSEI test set. The predicted sentiment intensity by our method is close to ground truth. 36

4.1 Accuracy scores for VCR test set. *VQA-GNN* outperforms *SGEITL+VLBERT* model on Q→AR metric by **3.2%**, and achieves competitive accuracy with SOTA methods, which have a close number of parameters but SOTA methods require a large amount of image caption data in pre-training process (over 13x larger than our model), *e.g.*, *"UNITER-L"*, *"ERNIE-ViL-B"*, *"RESERVE-B"*. Moreover, *"RESERVE-L+VQA-GNN"* outperforms *RESERVE-L* by **2.3%** on Q→AR metric. 49

4.2 All modules in the multimodal semantic graph help boost the final performance. Here, "scene-graph" includes node z and node p, "concept-graph" includes node z. 50

4.3 Ablation 1 and Ablation 2 indicate a single GNN on the multimodal semantic graph w/o and w/ direct cross-modal edges, respectively (Fig. 4.4). VQA-GNN with two modality-specialized GNNs on the multimodal semantic graph achieves the best score. 50

4.4 Accuracy scores on the GQA validation set. All models are trained under the realistic setup of not using the annotated semantic functional programs. 53

4.5 Ablation results on the effect of our proposed bidirectional fusion for GQA. 53

4.6 All models are trained only on the VCR dataset. Compared to the "SGEITL+VLBERT" model that inputs a visual scene graph to VL-BERT network, VQA-GNN applied to a well-structured multimodal semantic graph improves accuracy on Q→AR metric by over **4%**. . . 54

4.7 Comparison results on the different question types. VQA-GNN performs better than RESERVE-L for "Will", "Have" and "Can/Should" question types. 55

5.1 The size of training (Train), validation (Valid), and testing (Test) set for individual classes. 68

5.2 Samples from GAF 3.0 validation set. The description is annotated in the Japanese language and added to the training and validation dataset for training LDNN. A high score indicates strong cohesion among people shown in the image. 69
5.3 Comparison results of Japanese and English-based language encoder. 70
5.4 Comparison results of fine-tuned DNNs and visual LDNNs on validation(Val) and test sets. 70
5.5 Results of comparing ResNet50-based visual LDNN and state-of-the-art methods. 71
5.6 Comparison results of ablative study 1. "AL" and "KL" denote the AdaptLayer and KnowledgeLayer, respectively. 72
5.7 Comparison results of ablative study 2. 73

6.1 The size of training (Train), validation (Valid), and testing (Test) sets. 84
6.2 A video frame randomly selected from the VGAF validation set, the bounding boxes on the image indicate the region of objects. We added the video situation description in the Japanese language and used it as implicit information for training. 84
6.3 Comparison results of unimodal models on the validation set. 86
6.4 Comparison results with a baseline on the validation set. 86
6.5 Submission accuracy for the testing set. 89

7.1 Dataset size. MOSEI uses the same dataset as [1]. 101
7.2 The hyperparameters for training VideoAdviser. Here, "B" denotes the batch size, "Audio logit" denotes the output of the audio encoder for VEGAS (see Fig. 7.4). 104
7.3 Comparison results for MOSI and MOSEI. Our model reduces the state-of-the-art UniMSE's MAE score by **12.3%** for MOSI, and VAE-AMDT's MAE by **2.4%** for MOSEI. Here, (↓) indicates the lower the MAE, the better the performance, and (↑) indicates the vice-versa. (*) indicates the results produced on the modified MOSEI dataset. . . 105
7.4 The mAP comparison results with state-of-the-art models for VEGAS. Here, "V" and "A" indicate "Video" and "Audio", respectively. 106

7.5 Efficiency comparison. VideoAdviser is able to train a high efficiency-performance student model compared to state-of-the-art methods for inference. The student (RoBERTa-Base) outperforms the SOTA by over **3.0** point with nearly half the parameters. 107

7.6 Ablation results show the effects of components of the teacher model for multimodal knowledge distillation on MOSI dataset. 108

7.7 Effects in different student models. Our method improves the MAE score of pretrained language models by over **6.0** point on average. . . 108

7.8 Ablation results show the effects of step-distillation on audio and video modalities for VEGAS. Here, "w/ video distillation" indicates that the step-distillation is only adopted for the visual modality of the student model, "w/ audio distillation" indicates the other side, and "w/ audio and video distillation" indicates both sides (see Fig. 7.4). 109

7.9 Results of VideoAdviser trained with a combination of MOSI and MOSEI datasets. The model performs much better for both the MOSI and MOSEI test sets. Here, (*) denotes the result of the model trained on the individual dataset. 109

7.10 Ablation results show the effects of the proposed step-distillation loss for MOSI. 110

7.11 Comparison results between widely-known knowledge distillation loss and the proposed step-distillation loss for VEGAS. 110

7.12 ASO scores of models with different distillation objectives studied in Sec. 7.5.6. For "VideoAdviser (ours) → baseline", $\epsilon_{\text{min}} = 0$ indicates that VideoAdviser (ours) consistently outperform baseline. Here, the baseline denotes VideoAdviser (ours) w/o step-distillation. 111

List of Figures

2.1 A conceptual figure demonstrates that humans pay attention to representative emotional information from visual and audio modalities without considering multimodal alignment issues. 10

2.2 Overview of Multi-Attention Fusion Network. MAFN consists of two types of attention mechanisms: (1) **intra-modality attention mechanism** is computed to dynamically highlight central features of an unimodal video frame sequence; (2) **Inter-modality attention mechanism** is computed to automatically highlight specific modal features from giving multiple modalities. 12

2.3 Intra-modality attention mechanism. ⊗ denotes the element-wise product and ⊕ denotes the sum of the element-wise product for each time step. 14

2.4 Inter-modality attention mechanism. ⊗ is the element-wise product and ⊕ represents the sum of the element-wise product of all modalities. 15

2.5 Confusion matrix of the fusion result of MAFN w/ multimodal domain adaptation module shown in Tab. 2.1. 18

3.1 A conceptual diagram illustrates the distribution of various modalities in diversity. VAE-AMDT is designed to transfer unimodal representations to a joint sentiment embedding space. As a result, we obtain discriminative sentiment multimodal representations and make it easier to predict sentiment intensity. "△" and "◯" indicate "non-negative" and "negative" respectively. 21

3.2 Overview of our method: we first perform self-attention (§3.5.1) and cross-attention (§3.5.1) using preprocessed sequence features V, L and A, and then we perform VAE-AMDT that consisting of three VAEs and two generators G and one discriminator D to reduce distance difference between unimodal representations (§3.5.2). Finally, we use the encoded unimodal representations as the input of triple-attention (§3.5.1) to output one sentiment intensity result. Here, unimodal representations x_v, x_l, and x_a indicate concatenations of the output of attention layers for each modality. μ_v, μ_l and μ_a are encoded unimodal representations with VAE-AMDT. 25

3.3 Annotation distributions on (a) MOSI and (b) MOSEI. We show "negative" classes in red color and "non-negative" classes in blue color. 29

3.4 The precision-recall curve is created by using VAE-AMDT's test prediction results on both datasets. The curve indicates that VAE-AMDT outperforms Self-MM when both precision and recall scores exceed 0.8. Here, a better model should perform better for both metrics. 33

3.5 Visualization result on MOSEI. The green color indicates the "negative" class and the red color indicates the "positive" (including "neutral") class. The model (w/ VAE-AMDT) classifies both classes by discriminative representations. 34

4.1 Overview of *VQA-GNN*. Given an image and QA sentence, we obtain unstructured knowledge (*e.g.*, QA-concept node p and QA-context node z) and structured knowledge (*e.g.*, scene-graph and concept-graph), and then unify them to perform bidirectional fusion for visual question answering. 39

4.2 Reasoning procedure of VQA-GNN. We first build a multimodal semantic graph for each given image-QA pair to unify unstructured (*e.g.*, "node p" and "node z") and structured (*e.g.*, "scene-graph" and "concept-graph") multimodal knowledge (§4.4.1). Then we perform inter-modal message passing with a multimodal GNN-based bidirectional fusion method (§4.4.2) to update the representations of node z, p, v_i and c_i for $k + 1$ iterations in two steps. Finally, we predict the answer with these updated various node representations (§4.4.3). Here, "S" and "C" indicate scene-graph and concept-graph respectively. "LM_encoder" indicates a language model used to finetune QA-context node representation, and "GNN" indicates a relation-graph neural network for iterative message passing. 42

4.3 The process of concept-graph retrieval involves the calculation of similarity between concept-graph nodes and the answer context, denoted as $Relev(e|a)$. 44

4.4 Ablation architectures. We find that our final VQA-GNN architecture with two modality-specialized GNNs overcomes the representation gaps between modalities (§4.4.1). 51

4.5 Illustration of two knowledge fusion methods: our proposed bidirectional fusion v.s. the unidirectional fusion baseline. 54

4.6 Interpreting *VQA-GNN*'s reasoning process across multimodal knowledge domains by indicating attention weight of the relationship between nodes. Arrows indicate the direction of the relationship, and darker and thicker edges indicate higher attention weights. The red color highlights the message passing routine for reasoning the correct answer and the gray color indicates the opposite. 56

5.1 A conceptual diagram illustrating the injection of linguistic knowledge into a visual prediction model. 60

5.2 Architecture for training LDNN. The visual model and the language model are jointly trained to inject linguistic knowledge into the visual model. The visual encoder is a pretrained DNN model for extracting visual representations. The Language encoder is a BERT-based pretrained model for extracting language representations. The bottom part of this figure shows the learning process of injecting linguistic knowledge. 64

5.3 LDNN visual model used for predicting the cohesive level of an image. 65

5.4 A knowledgeLayer ablation network for ablative Study 1. 72

5.5 Comparison results of LDNN and DNN baselines on the testing set. . 74

6.1 A conceptual figure demonstrating the process of injecting implicit knowledge into an audiovisual model, and the emotion prediction by using only explicit features. 76

6.2 K-injection audiovisual network: An end-to-end architecture for training an audiovisual model in which the implicit knowledge representing the emotions of a group is transferred from two K-injection subnetworks. The top part shows the audiovisual subnetwork. “FF”, “Norm”, and “Sum” denote the feedforward layer, normalization method, and attention-weighted sum calculator, respectively. “Attention” indicates multi-head self attention and “Cross” indicates multi-head cross attention. The bottom part shows the linguistic K-injection subnetwork and acoustic K-injection subnetwork. “GRU” denotes the gated recurrent unit layer, and “Self” indicates self-attention. The symbol of “$\oplus$” presents a residual connection. 80

6.3 Training progress of the cross-multi head attention based audiovisual model 87

6.4 Training progress of the K-injection subnetwork-based audiovisual model. 88

7.1 A conceptual diagram illustrates the difference between the conventional system and our system: our system focuses on transferring multimodal knowledge from a multimodal fundamental model (*e.g.*, CLIP) to a language fundamental model (*e.g.*, RoBERTa-Large), and requires text only to achieve high efficiency-performance inference. On the other hand, the conventional system focuses on multimodal fusion and requires complex modules (diverse modal encoders and a multimodal fusion module) for inference. 92

7.2 Architecture of VideoAdviser using a CLIP-based model (the teacher) to distill multimodal knowledge of video-enhanced prompts to a RoBERTa-based model (the student): the teacher model utilizes pretrained CLIP's text and visual encoders, and a facial expression encoder to obtain the sentiment class text embedding, the frame-level embedding, and the facial expression embedding. Then, the teacher model employs CCT, MIT, MLP, and a video-specific prompting module, and minimizes a binary sentiment classification loss and a sentiment regression loss. Meanwhile, the student model is finetuned on speech text by minimizing a sentiment regression loss and a step-distillation loss (the region in purple). During inference, the speech text is used to enable sentiment intensity prediction. Here, CCT, MIT, and MLP stand for the cross-frame communication transformer, multi-frame integration transformer, and multi-layer perceptron, respectively. . . . 97

7.3 Label distribution of (a) MOSI and (b) MOSEI. The synthesized binary classification label is illustrated in different colors (the "negative" class in red color and the "non-negative" class in blue color). . . 101

7.4 Architecture of VideoAdviser for audio-visual retrieval task using a CLIP-based model (the teacher) to distill multimodal knowledge of video-enhanced prompts to an EICS-based audio-visual model (the student). The teacher model is finetuned for the audio event classification to distill multimodal knowledge to the student model via the step-distillation loss (the region in purple). We adopt 3-layer MLP with 128-dimensional hidden layers. 103

7.5 Visualization of logistic knowledge distribution with and without the step-distillation objective loss. The top row plots the histograms of logit by applying the step-distillation, and the bottom row indicates the vice-versa. The groudTruth indicates the label distribution, and text_logit indicates the predicted regression score of the student model. Our method using the step-distillation (the top) demonstrates a distribution of regression scores close to the groundTruth, affected by the knowledge distribution of the "video_logit" and "video_enhanced_prompt_logit". 111

Chapter 1

Introduction

1.1 Motivation

Humans effortlessly interpret the real world around them by utilizing multiple modalities such as vision, text, and audio. Mimicking human intelligence to integrate multimodal information has demonstrated remarkable success across a range of tasks. (1) video-level sentiment analysis task [2, 3] aims to make sentiment predictions from unified diverse modal representations encoded by readily available pretraining models by utilizing diverse modal representations derived from readily available pretraining models [4, 5, 6]. (2) Cross-modal retrieval tasks include text-visual retrieval [7, 8], as well as audio-visual retrieval [9, 10, 11]). These tasks aim to retrieve one modal content from a pool of candidate belongings to another modality by learning their cross-modal representations within a common domain space. (3) Visual question answering (VQA) task aims to provide answers to questions about a visual scene [12]. It is crucial in many real-world tasks including scene understanding, autonomous vehicles, search engines, and recommendation systems [13, 14, 15].

To obtain holistic features of diverse modalities for effective fusion, preliminary attempts were made by leveraging deep learning techniques. For instance, convolutional neural networks (CNN) such as VGG [16] and ResNet [17] have broadly been used for extracting visual features. Convo lutional recurrent neural networks (CRNNs) provide a robust approach to audio feature extraction [18]. Word2Vec is a widely used method for learning word embeddings, which are dense vector representations of words in a continuous vector space [19]. Concatenating diverse modal features encoded with these techniques within a single vector enables the fusion of multiple modalities. However, the dynamic interaction across modalities can not be captured for modeling expressive multimodal representations. Consequently, early deep learning model-based methods give limited performance improvement [20].

Recently, the Transformer model built with attention mechanisms has emerged as a promising technique of deep learning models, which allows models to selectively process information, enhancing their ability to dynamically capture important features and relationships within the data [21]. Indeed, the Transformer model pretrained on large amounts of data has achieved remarkable success by capturing potential features from raw data, such as visual models (*e.g.*, ViT [4]), language models (*e.g.*, RoBERTa [5] and GPT-3 [22]), and audio models (*e.g.*, VGGish [6], and SoundNet [23]). As a result, there has been increasing interest in multimodal fusion on unifying multimodal features encoded by these pretrained transformer models. Moreover, recent visual-language pretraining (VLP) models [24, 25, 26, 27] employ the pretrain-and-finetune approach, where they train a multimodal transformer model on large-scale visual-language datasets, and then finetune the pretrained model on the downstream VQA datasets, *e.g.*, RESERVE-L model [28] is pretrained using 1 billion image-caption data including video frames, text, and audio. VLP models have shown strong performance by integrating various modalities and capturing interactions among them using a large number of parameters. Thereby, these models are gaining attention for their potential to serve as foundational multimodal models capable of addressing a wide range of tasks [29, 30].

However, while VLP models enable the learning of potential multimodal representations, the presence of various distributions among modalities yields inconsistent fusion representations. As a result, it prevents systems from learning discriminative multimodal representations. For instance, consider the multimodal sensitive analysis case where a multimodal representation is formed by concatenating visual and language features from separate domains. Distributional differences between these representations can make it difficult for the model to accurately associate this combined representation with a consistently positive emotional state. Moreover, existing multimodal systems are typically built on the assumption that all modalities exist, and the lack of modalities always leads to poor inference performance. These limitations prevent enhancing the performance efficiency of real-world multimodal applications. In addition, multimodal systems require tremendous parameters, making inference computationally inefficient. Therefore, studying how to improve the performance efficiency of multimodal representation learning is crucial to achieving realistic multimodal systems.

1.2 Thesis Outline

This thesis involves various endeavors aimed at improving the performance efficiency of multimodal representation learning. These efforts not only facilitate high-performance multimodal fusion but also enhance the efficiency of multimodal inference—an unimodal system can achieve competitive performance with a multimodal system even when provided with a specific unimodal input. This thesis consists of two parts.

Part 1: Improving high-performance multimodal fusion. We focus on improving high-performance multimodal fusion by converting devise modalities into common representation distribution domains.

- In Chapter 2, we propose a multiple attention fusion network (MAFN), which includes a multimodal domain adaptation module designed to obtain expressive multimodal representations capturing dynamic interactions across modalities.

- In Chapter 3, we also propose a VAE-based adversarial multimodal domain transfer (VAE-AMDT) method to further improve effective multimodal fusion by reducing the distribution difference between diverse modal representations.

- In Chapter 4, we propose a new visual question answering (VQA) method called VQA-GNN that performs bidirectional fusion between unstructured and structured multimodal knowledge through a language model (LM) to eliminate fusing distinct modalities directly.

Part 2: Achieving effective multimodal transfer learning. We focus on achieving effective multimodal transfer learning aiming for efficient multimodal systems.

- In Chapter 5, we propose a language knowledge injectable deep neural network (LDNN) to achieve a visual model enhanced with implicit language knowledge encoded by pretrained LM.

- In Chapter 6, we extend LDNN proposed in chapter 5 to propose an end-to-end architecture called implicit knowledge injectable cross attention audiovisual deep neural network (K-injection audiovisual network), aiming to transfer the potential knowledge provided by pretrained language and audio models into an audiovisual model built from raw visual and audio data.

- In Chapter 7, we develop VideoAdviser, a video knowledge distillation method to transfer multimodal knowledge from a multimodal fundamental model (*e.g.*, CLIP) to a specific modal fundamental model via optimizing a step-distillation objective loss.

We conclude and discuss future challenges in Chapter 8.

1.3 Contributions

The contributions of this thesis are summarized as follows:

- We proposed a VAE-based adversarial multimodal domain transfer learning method. By jointly training it with a multi-attention module, our method balanced the modal distribution difference between any modality pairs and reduced their average distance in total. As a result, we obtained discriminative multimodal representations to further improve the performance of multimodal tasks (*e.g.*, video-level sentiment analysis).

- We proposed VQA-GNN to perform bidirectional fusion to unify multimodal knowledge via graph neural networks for expressive concept-level reasoning. Compared with existing works, which only perform a late fusion or unidirectional fusion from unstructured knowledge to structured knowledge, our method makes two technical innovations: **bidirectional fusion** and **multimodal GNN**. By fairly comparing with existing works, our method substantially outperforms existing models. Ablative studies further suggest the efficacy of the bidirectional fusion and multimodal GNN method in unifying unstructured and structured multimodal knowledge.

- We proposed a novel multimodal knowledge distillation method, VideoAdviser, which leverages the strengths of learned multimodal space of the CLIP-based teacher model and large-scale parameters of the RoBERTa-based student model to perform multimodal knowledge transfer by optimizing a step-distillation objective loss. By comparing to state-of-the-art methods (SoTA), our method significantly outperforms SoTA methods with a single modal encoder used in inference, suggesting its strengths in high performance efficiency. In particular, the comparison results demonstrate the efficacy of our proposed step-distillation objective loss in improving multimodal knowledge distillation to achieve a modality-agnostic multimodal system.

1.4 Bibliographic Remarks

The works presented in this thesis are based on the following publications:

- Chapter 2: **Yanan Wang**, Jianming Wu, Keiichiro Hoashi. *Multi-attention fusion network for video-based emotion recognition.* International Conference on Multimodal Interaction (ICMI), 2019.

- Chapter 3: **Yanan Wang**, Jianming Wu, Kazuaki Furumai, Shinya Wada, Satoshi Kurihara. *VAE-Based Adversarial Multimodal Domain Transfer for Video-Level Sentiment Analysis.* IEEE Access, 2022.

- Chapter 4: **Yanan Wang**, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, Jure Leskovec. *VQA-GNN: Reasoning with Multimodal Knowledge via Graph Neural Networks for Visual Question Answering.* International Conference on Computer Vision (ICCV), 2023.

- Chapter 5: **Yanan Wang**, Jianming Wu, Jinfa Huang, Gen Hattori, Yasuhiro Takishima, Shinya Wada, Rui Kimura, Jie Chen, Satoshi Kurihara. *LDNN: Linguistic knowledge injectable deep neural network for group cohesiveness understanding.* International Conference on Multimodal Interaction (ICMI), 2020.

- Chapter 6: **Yanan Wang**, Jianming Wu, Panikos Heracleous, Shinya Wada, Rui Kimura, Satoshi Kurihara. *Implicit knowledge injectable cross attention audiovisual model for group emotion recognition.* International Conference on Multimodal Interaction (ICMI), 2020.

- Chapter 7: **Yanan Wang**, Donghuo Zeng, Shinya Wada, Satoshi Kurihara. *VideoAdviser: Video Knowledge Distillation for Multimodal Transfer Learning.* IEEE Access, 2023.

Part 1: Improving high-performance multimodal fusion

Chapter 2

Multi-Attention Fusion Network for Video-based Emotion Recognition

In this chapter, with the goal of capturing dynamic multimodal interactions for improving multimodal fusion, we propose a multiple attention fusion network (MAFN) involving a multimodal domain adaptation module. MAFN consists of two types of attention mechanisms: (1) **intra-modality attention mechanism** is computed to dynamically highlight central features of an unimodal video frame sequence; (2) **Inter-modality attention mechanism** is computed to automatically highlight specific modal features from giving multiple modalities. **Multimodal domain adaptation module** is further employed to reduce the distance of processed unimodal representation by (intra and inter)-modality attention mechanisms to enhance expressive multimodal fusion.

2.1 Introduction

Emotion is an important part of human communication, and human-like automatic emotion recognition (AER) technology is essential to achieve communication between humans and artificial intelligence (AI). With the development of deep learning technology, there are many studies on identifying human emotions through processing facial expression, audio, and language information [31, 32, 33, 34, 35]. However, humans routinely recognize emotions by integrating multimodal information, and in particular, both visual and audio information is very important for

AER [36, 37, 38, 39, 40, 41]. Therefore, modeling human emotion recognition is pivotal for obtaining effective emotional information from multimodal data.

Recently, most of the research on multimodal emotion recognition tasks focuses on the aspects of extracting representative modality features and defining dynamic interactions between multiple modalities [36, 37, 38]. It has become easier to extract representative modality features utilizing deep learning technology. For instance, fine-tuning based on state-of-the-art CNNs (AlexNet [42], VGG [16], ResNet [17], SENets [43]) is very useful for capturing the fine-grained facial expression features, and Long Short-Term Memory units (LSTMs) are another deep learning technology that can be used to store information including short-term interaction of time-step features in memory over time. Based on these deep learning technologies, there are many works on defining multimodal dynamic interactions by associating a relevance score with each LSTM memory unit [40, 41]. To implement these works, we need to align not only the dimensions of the multimodal features but also the sequence lengths of all modalities, such as by duplicating previous frame features. However, such a forced alignment strategy not only adversely affects the extraction of important modality features, but also loses the opportunity to establish the optimal interaction between multimodal information. As we know, humans recognize emotion through combining complex multimodal information and tend to pay attention only to important information across different modalities. For example, some people always speak while keeping a smile, and others may speak loud but not angry. Therefore, we consider that humans do not recognize emotions based on alignment between modalities. Fig. 2.1 shows an example where humans extract only important non-aligned modality information to recognize emotions.

In this chapter, we propose a multiple attention fusion network (MAFN) aiming to improve emotion recognition performance by modeling human emotion recognition mechanisms. We implement MAFN utilizing the attention mechanism which is designed to aggregate essential information over time corresponding to the results [44]. The attention mechanism has been successfully applied to LSTM for highlighting the most important sequence features and also has been used in an attempt to build a connection between visual and audio modalities [45, 46]. MAFN consists of two types of attention mechanisms: (1) **intra-modality attention mechanism** is computed to dynamically highlight central features of an unimodal video frame sequence; (2) **Inter-modality attention mechanism** is computed to automatically highlight specific modal features from giving multiple modalities. Moreover, reduce

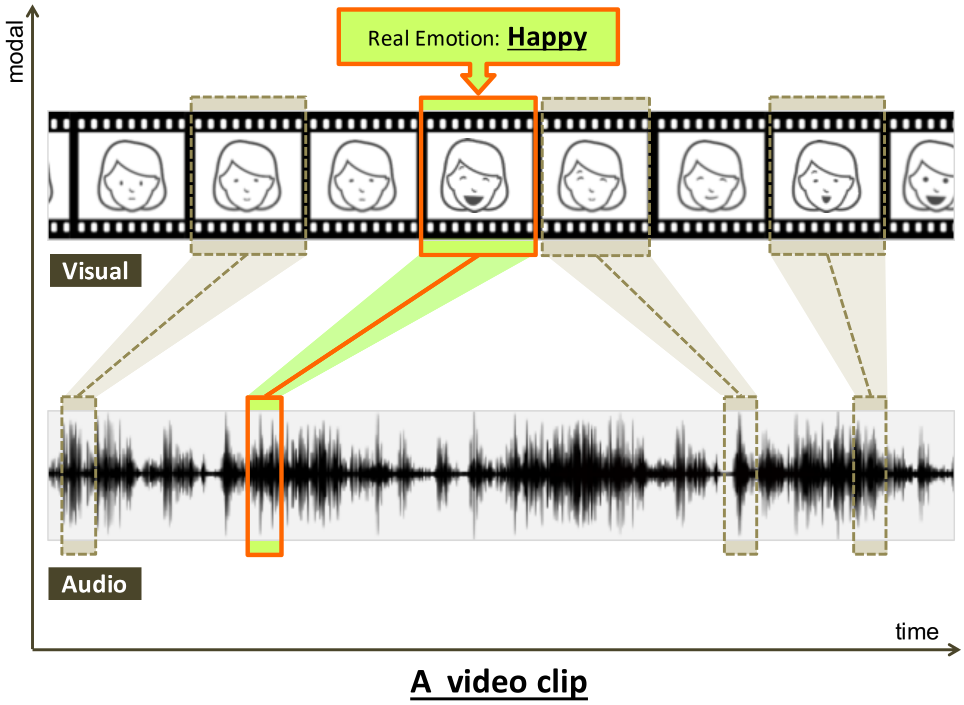


Figure 2.1: A conceptual figure demonstrates that humans pay attention to representative emotional information from visual and audio modalities without considering multimodal alignment issues.

the distance of processed unimodal representation by (intra and inter)-modality attention mechanisms to enhance expressive multimodal fusion. We do not apply a forced alignment strategy to MAFN, which aims to extract the meaningful features of each modality and to establish the optimal interactions between multimodal information.

We first demonstrate MAFN on the AFEW dataset [36, 47] which contains two types of modality information: visual and audio modalities. As a result, MAFN achieves **58.65%** recognition accuracy with the AFEW testing set, which is a significant improvement compared with the baseline of 41.07% [36].

2.2 Related Works

In terms of human emotion understanding, 93% rely on nonverbal (facial expressions: 55%, audio: 38%), and 7% rely on verbal language [48]. This is why there are many works focused on facial expression recognition (FER) and audio emotion recognition

(AER) tasks [32, 49, 31, 33, 34]. Most of these works utilize deep learning technology to extract informatic features for achieving high emotion recognition [32, 31, 34]. For instance, Tang [32] built a DNN-based structure that combines convolutional networks with L2-SVMs as an activation function and won the ICML 2013 FER challenge. This work [31] employed and confirmed that modern DNN architectures (VGG, ResNet, Inception [50]) have the potential to extract fine-grained facial expression features to improve FER performance. On the other hand, in terms of AER tasks, the commonly used features include pitch, log-Mel filter banks energies (log-Mels), and Mel-frequency cepstral coefficients (MFCCs) [33, 34]. This work [51] employed four types of convolutional operation to extract more comprehensive emotion features with log-Mels features.

Due to human emotions being complex and diverse, some emotions are difficult to recognize by a single modal signal. Recently, there have been many research works that focus on combining multimodal features to improve emotion recognition performance. Audio-video emotion challenge is a part of Emotion Recognition in the Wild (EmotiW) [36], which is a series of benchmarking works focusing on affective computing issues. This work [37] proposed a method that concatenates facial and audio features extracted through several DNN models and achieves the best accuracy in the Acted Facial Expression in the Wild (AFEW) testing set of 61.87%. This work [38] proposed a multiple spatio-temporal feature fusion (MSFF) framework. They fine-tuned a pretrained model using facial expression images to extract facial expression features and then applied VGG-19 and BLSTM models for extracting audio emotion features. They finally employed a decision fusion approach to enhance the performance of emotion recognition. However, these methods did not consider the interactions between different modalities. In contrast, the work [40] considered the consistency and properties complementary of different modal information and proposed a memory fusion network that models modal-specific and cross-modal interactions over time to effectively capture emotion features and achieved high performance on the CMU-MOSI dataset [39]. Furthermore, the work [41] proposed a Dynamic Fusion Graph neural model, which aims to model diverse multimodal interactions such as unimodal, bimodal, and trimodal interactions. As a result, it can dynamically alter multimodal features based on the importance of individual multimodal dynamics during fusion. While [39] and [41] can capture interactions of different modalities dynamically, it is required to align different modalities by taking the average of their modal features over the word utterance time interval.

However, the word-based alignment strategy may miss the opportunity to capture more effective interactions between modalities.

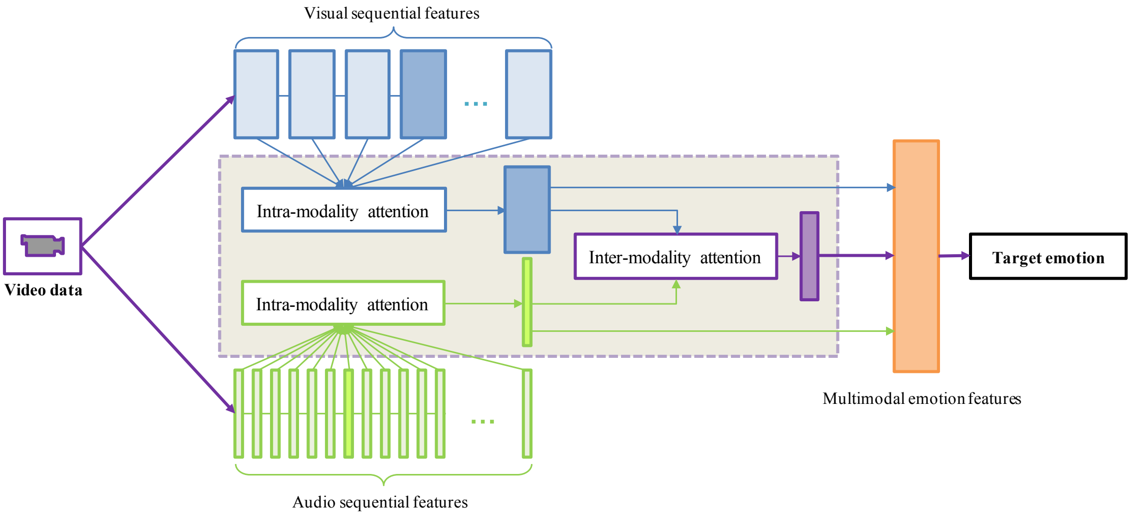


Figure 2.2: Overview of Multi-Attention Fusion Network. MAFN consists of two types of attention mechanisms: (1) **intra-modality attention mechanism** is computed to dynamically highlight central features of an unimodal video frame sequence; (2) **Inter-modality attention mechanism** is computed to automatically highlight specific modal features from giving multiple modalities.

2.3 Multi-Attention Fusion Network (MAFN)

In this chapter, we propose a method based on the concept that humans tend to pay attention to important information across different modalities even if the modality information is not aligned with each other. Instead of an alignment strategy, our proposed method applies multiple attention mechanisms that aim to model intra/inter-modal dynamics to capture more meaningful multimodal emotion features compared to previous works.

Fig. 2.2 shows an overview of our proposed MAFN. MAFN consists of two types of attention mechanisms: 1) The intra-modality attention mechanism takes as input single-modal features embedded by bi-directional LSTM, and dynamically outputs meaningful modal attentional features corresponding to the target emotion. 2) The inter-modality attention mechanism takes as input modal attentional features extracted from the intra-modality attention mechanism, and outputs inter-modality attentional features based on modality importance. MAFN finally concatenates visual, audio, and inter-modality attentional features to generate multimodal emotion features. MAFN also applies the multimodal domain adaptation module to increase the effectiveness of the inter-modality attention mechanism. The following three subsections describe the attention mechanisms and multimodal domain adaptation module in detail.

2.3.1 Intra-modality attention mechanism

As mentioned in sec. 2.1, attention mechanisms [45, 21, 52] are designed to focus on certain aspects of sequence data and aggregate important information over time to correspond to the target results. The attention weight is not constrained by the trained model and is adaptively calculated based on the importance of the sequence data to extract necessary features [21, 52]. We apply the self-attention mechanism [44] to both visual and audio sequence data to extract the important visual and audio emotion features, respectively.

As shown in Fig. 2.3, giving a sequential feature $X^m = [x_t^m : t \leq T, x_t^m \in \mathbb{R}^{d_x^m}]$ for the $m \in \{v, a\}$ modality, the intra-modality attention mechanism employs 2-layer BLSTM followed by the intra-modality attention layer to generate m-modal attentional feature $\hat{x}^m$. Here, T and d_x^m denote the sequence length and the dimension size respectively. Following Eqs. 2.1 and 2.2, we first calculate the output of the 2-layer $H^m = [h_t^m : t \leq T, h_t^m \in \mathbb{R}^{d_h^m}]$, and then calculate the modal attention weight $A^m = [a_t^m : t \leq T, a_t^m \in \mathbb{R}^1]$ by introducing two weight matrix W_{s1} and W_{s2} with shapes of $(d_h^m, d_h^m/2)$ and $(d_h^m/2, 1)$. We use the activation function ReLU with a range of $[0, infinity]$ and the softmax function to ensure all weights sum up to 1. Finally, we follow Eq. 2.3 to calculate m-modal attentional feature $\hat{x}^m$.

$$H^m = \mathrm{BLSTM}(\mathrm{BLSTM}(X^m)) \tag{2.1}$$

$$A^m = \mathrm{softmax}\left(W_{s2}\,\mathrm{ReLU}(W_{s1}H^m)\right) \tag{2.2}$$

$$\hat{x}^m = \frac{\sum_{t=0}^{T}(a_t^m \cdot H^m)}{T} \tag{2.3}$$

2.3.2 Inter-modality attention mechanism

Considering that humans recognize emotions by instinctively combining important emotion features from different modalities, we introduce the inter-modality attention mechanism to reveal and unify the dynamic interaction between visual and audio modalities. As shown in Fig. 2.4, we have visual-modal attentional feature $\hat{x}^v$ and

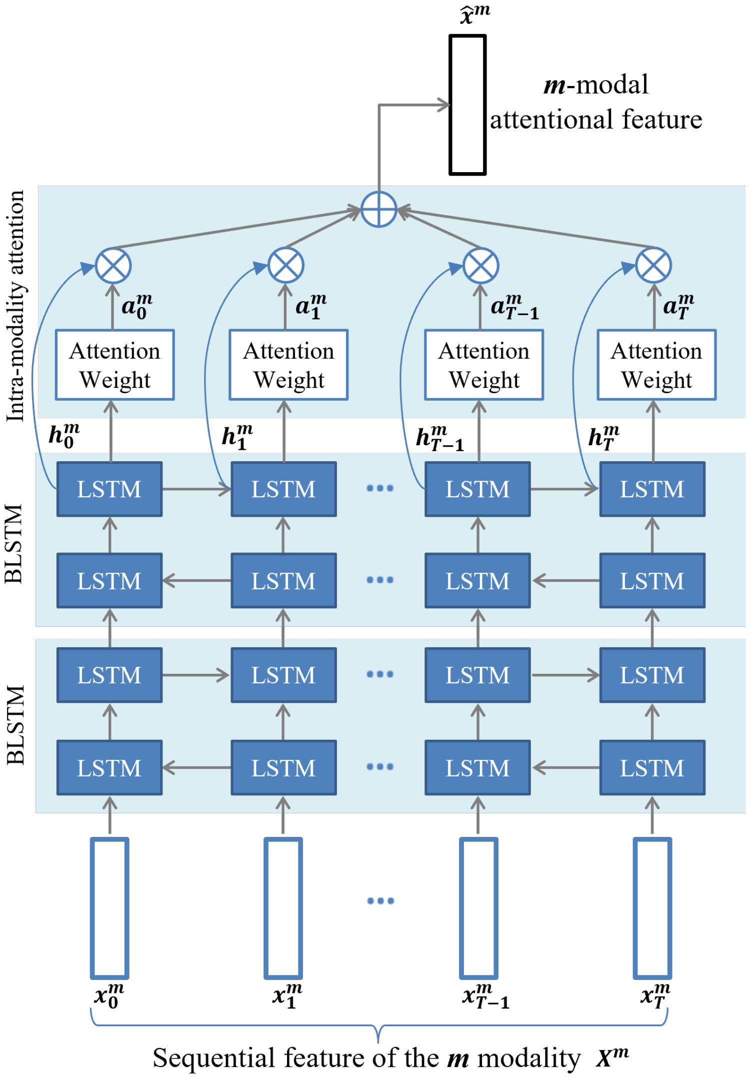


Figure 2.3: Intra-modality attention mechanism. ⊗ denotes the element-wise product and ⊕ denotes the sum of the element-wise product for each time step.

audio-modal attentional feature $\hat{x}^a$ for the input of the inter-modality attention mechanism. We first define $Z = [W_v\hat{x}^v, W_a\hat{x}^a]$ where W_v and W_a are weight matrix with shapes of (d_h^v, d_m) and (d_h^a, d_m) respectively. d_h^v and d_h^a are the dimension size of the visual and audio modality attentional features, and d_m is a fixed size for generating inter-modality attentional features. Following Eq. 2.4, we calculate attention weight $A^{v-a} = [a_0^{v-a}, a_1^{v-a}]$. The inter-modality attentional feature $\hat{x}^{v-a}$ is computed following Eq. 2.5:

$$A^{v-a} = \text{softmax}\left(W_{m2}\,\text{ReLU}(W_{m1}Z)\right) \tag{2.4}$$

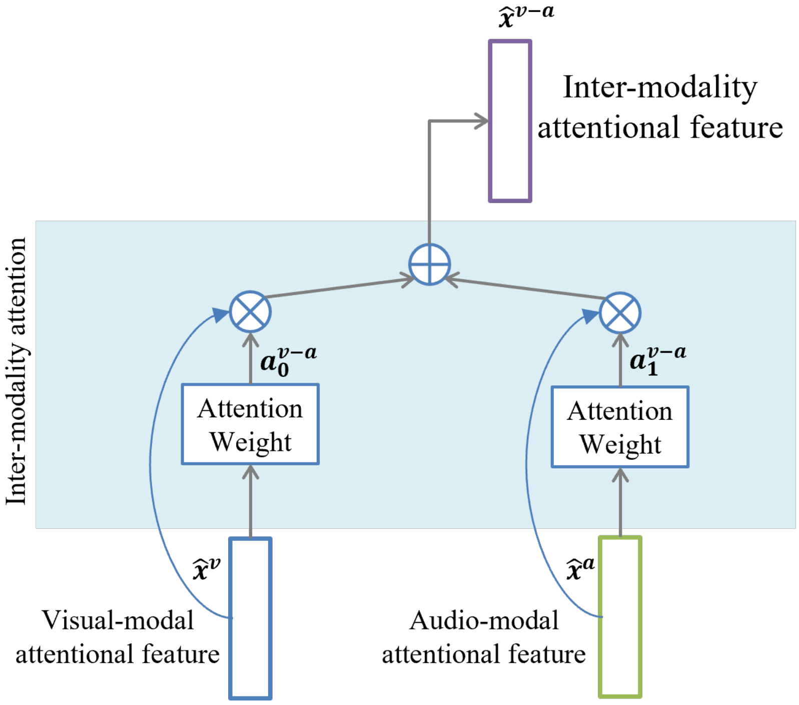


Figure 2.4: Inter-modality attention mechanism. $\otimes$ is the element-wise product and $\oplus$ represents the sum of the element-wise product of all modalities.

Here, W_{m1} and W_{m2} are weight matrix with different shape of $(d_m, d_m/2)$ and $(d_m/2, 1)$.

$$\hat{x}^{v-a} = \frac{a_0^{v-a} \cdot Z[0] + a_1^{v-a} \cdot Z[1]}{2} \tag{2.5}$$

2.3.3 Multimodal domain adaptation module

To make MAFN learn impressive interactions between visual and audio attentional features, we introduce the distLoss to force MAFN to learn the mappings between different modal feature domains by minimizing the distance between visual and audio modal attentional features. Along with the target training for emotion classification, the MAFN additionally optimizes the distLoss with mean squared error (MSE) loss function as follows:

$$\text{distLoss}_{\text{v-a}} = \text{MSE}(Z[0], Z[1]) \tag{2.6}$$

where Z[0] and Z[1] represent the weighted visual and audio modal attentional features respectively.

2.4 Experiment

In this section, we describe the details of the experiment and evaluate the performance of MAFN on the AFEW dataset [36].

2.4.1 Data preprocessing

The AFEW dataset is collected from movies and TV shows and contains 773, 383, and 653 video clips in training, validation, and testing sets, respectively. The AFEW dataset is not large enough to be used to train representative emotion features from raw video clips. Therefore, the extraction of features that well represent emotions is the key to improving emotion recognition performance.

Visual features: To extract meaningful facial expression features as the input visual modal features to MAFN, we first convert video clips into image sequences at 16fps, then utilize OpenFace to detect the face in images, and finally apply the detected face images to 5 types of pretrained FER models to extract facial expression features with 8 dimensions, respectively. These pretrained FER models contain AlexNet [42], VGG_M [16], VGG_VD [53], SENet50 [43] and ResNet50 [17]. AlexNet, VGG_M and VGG_VD are fine-tuned on the Fer2013 dataset [54], and SENet50 and ResNet50 are fine-tuned on the Fer2013+ dataset [55]. Due to each video clip having a different sequence length, we process upsampling and downsampling to generate a fixed visual sequence length of 32 frames.

Audio features: We choose a basic set of audio features as the input audio modal features to MAFN. Firstly, we separate audio data from raw video data and then extract log-Mel filter banks with 64 dimensions after resampling audio data with a reduced sample rate from 48KHz to 44.1KHz. The audio frame length is 0.0415s, and the window length is 0.064s. Finally, we process upsampling and downsampling to generate a fixed audio sequence length of 128 frames.

Furthermore, we perform rolling feature engineering for both visual and audio features to generate new features that can capture temporal variational information in a video clip. We adopt three types of rolling methods (mean, standard deviation, and variation of max and min) to visual and audio features with two different rolling window sizes (visual features: 4, audio features: 6). As a result, the input visual sequential features are sized to 29-sequence-length by 48 dimensions, and the input audio sequential features are sized to 123-sequence-length by 192 dimensions.

2.4.2 Training setting

Considering the small size of training (773) and validation (383), we concatenate the training and validation sets and randomly re-split them at a rate of 90% for training and 10% for validation. In the training process, we perform the re-splitting process 5 times and choose the best model based on validation loss, then select the appropriate hyperparameter (dropout rate: 0.5, batch size: 128, learning rate: 0.0001, epochs: 200) to achieve the best performance of MAFN.

2.4.3 Result

We trained MAFN utilizing five types of visual features with log_Mel filter banks audio features and evaluated them on the re-splitting validation set. To further improve recognition accuracy, we assigned different weights depending on the validation accuracy of each model. The best result submitted is a fusion of the prediction results of the testing set based on the following weights: AlexNet: 0.5, VGG_M: 0.7, VGG_VD: 0.7, SENet50: 1.0, and ResNet50: 1.0.

As shown in Tab. 2.1, MAFN taking ResNet50 facial expressions features as visual features achieves the highest accuracy of 62.07% on the re-splitting validation set compared with other models. The fusion result is 58.65% on the testing set, which is a significant improvement compared with the baseline of 41.07% [36].

The confusion matrix of the fusion result on the testing set is shown in Fig. 2.5. MAFN achieves recognition accuracy by over 70% on angry, happy, and neutral emotion classes, while it cannot classify the disgust and surprise emotions correctly. We also observed that disgust and surprise emotions tend to be recognized as neutral emotions. We also submitted the results generated by MAFN without implementing the multimodal domain adaptation module and achieved 56.20% accuracy on the testing set, which is 2.45% lower than the fusion of MAFN models. The results of the re-splitting validation set are shown in Tab. 2.2. All scores are lower than MAFN implementing the multimodal domain adaptation module.

2.4.4 Discussion

According to the results of MAFN trained by utilizing different visual features, we observed that the variation in validation accuracy is so large that it is difficult to determine which type of visual features can achieve high performance on the testing

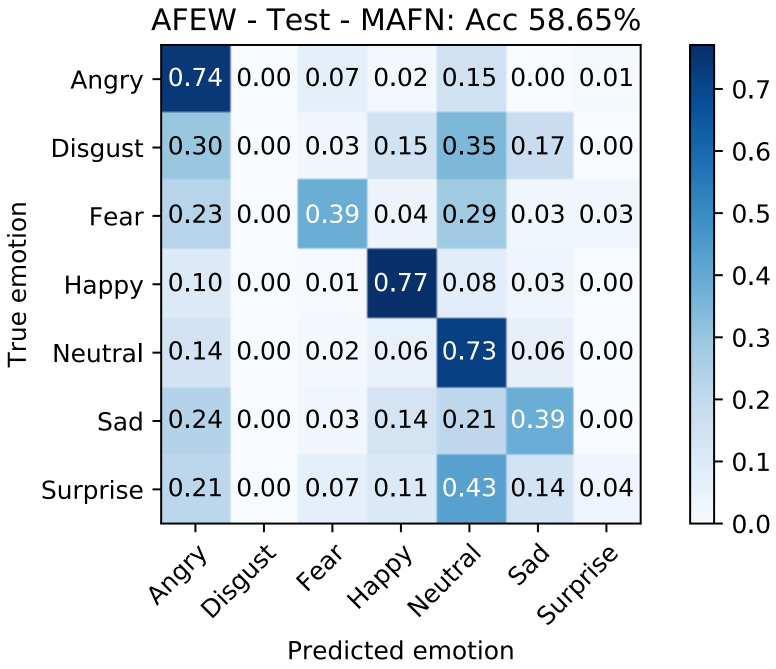


Figure 2.5: Confusion matrix of the fusion result of MAFN w/ multimodal domain adaptation module shown in Tab. 2.1.

set. Many previous works have attempted fusing as many different modal features or model types as possible to improve emotion recognition accuracy [36, 37, 38]. In contrast, we focused on new methods based on the analysis of human emotion recognition mechanisms utilizing basic modal features. The results show that MAFN is capable of improving emotion recognition accuracy. In addition, according to a comparison of the results shown in Tab. 2.1 and Tab. 2.2, we confirmed that the multimodal domain adaptation module is effective in improving emotion recognition performance, and consider that the multimodal domain adaptation module can accelerate the learning of interactions between modalities.

Meanwhile, similar to related works [37, 38], MAFN cannot classify disgust and surprise emotions, we consider that it is insufficient to classify them only based on visual and audio information extracted from the AFEW dataset, and the context information contained in the text should be analyzed together to capture more meaningful emotion features.

MAFN w/ multimodal domain adaptation module	Validation (%)	Test (%)
Visual encoder		
– AlexNet	50.00	-
– VGG_M	55.17	-
– VGG_VD	55.17	-
– SENet50	61.21	-
– ResNet50	62.07	-
Fusion model	-	**58.65**

Table 2.1: Comparison results of MAFN built using various visual encoders. Here, we use the modified validation and test datasets to evaluate the accuracy of the emotion recognition task.

MAFN w/o multimodal domain adaptation module	Validation (%)	Test (%)
Visual encoder		
– AlexNet	49.68	-
– VGG_M	51.30	-
– VGG_VD	52.17	-
– SENet50	58.26	-
– ResNet50	60.87	-
Fusion model	-	**56.20**

Table 2.2: Recognition accuracy of each MAFN w/o multimodal domain adaptation module on the modified dataset. The performance is reduced compared to the results in Tab. 2.1.

2.5 Conclusions

In this chapter, We proposed MAFN to model human emotion recognition mechanisms, which is based on the concept that humans pay attention to representative emotion information from visual and audio modalities. MAFN is constructed with intra and inter-modality attention mechanisms and a multimodal adaption module. It achieves competitive performance on the testing set without combining scores generated by other models.

Chapter 3

VAE-Based Adversarial Multimodal Domain Transfer for Video-Level Sentiment Analysis

In the last chapter, we present an approach to improve multimodal fusion by introducing a multimodal domain adaptation module to enhance the (intra and inter)-attention-based model for capturing multimodal interactions. In this chapter, to obtain more discriminative multimodal representations that can further improve systems' performance, we propose a VAE-based adversarial multimodal domain transfer (VAE-AMDT) and jointly train it with a multi-attention module to reduce the distance difference between unimodal representations.

3.1 Introduction

Video-level sentiment analysis is a task to predict people's sentiment intensity with a given video clip. It is an essential task for achieving high-level artificial intelligence (AI), and is expected to be applied to dialogue agents, virtual reality and social robotics, and so on [56]. To let AI systems have a better understanding of people's sentiment, existing methods fuse multimodal representations obtained from video frames (image), text, and audio, and predict sentiment intensity by doing regression analysis [40, 57]. How to obtain discriminative multimodal representations that can capture differences in sentiments across various modalities is a core issue for video-level sentiment analysis [20, 58, 59]. However, due to diverse distributions of various modalities (*e.g.*, one same sentiment intensity corresponds to different

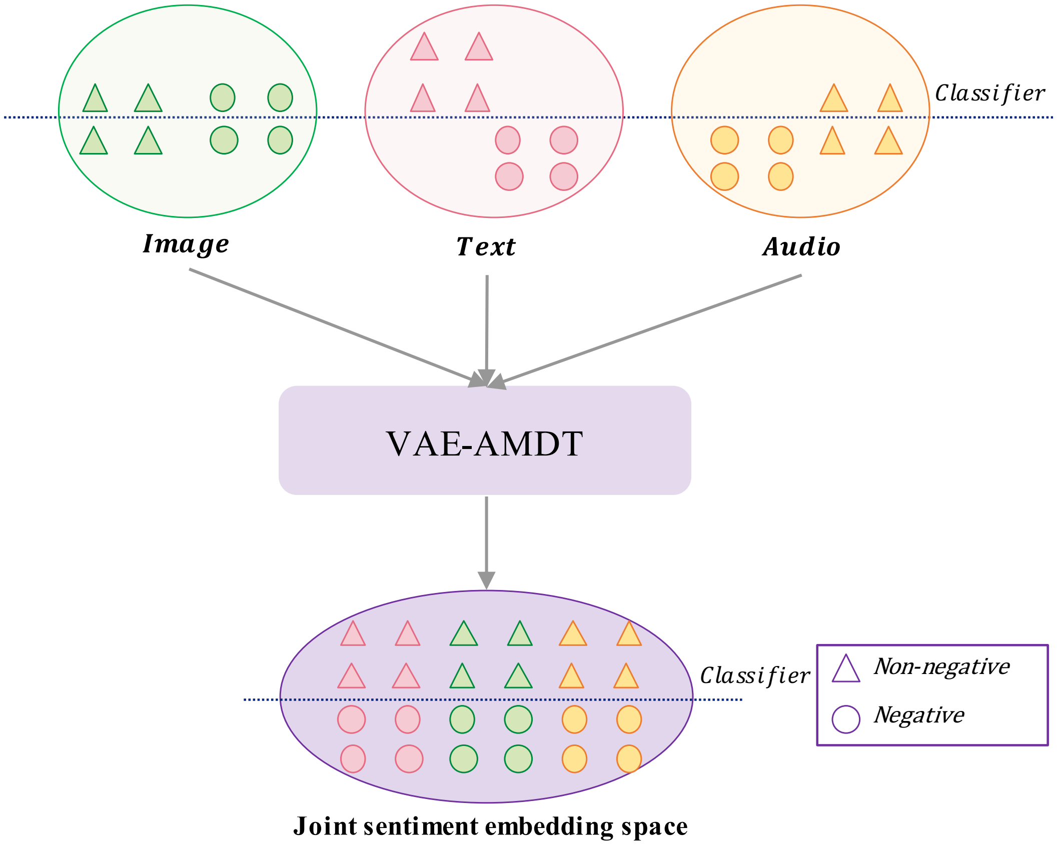


Figure 3.1: A conceptual diagram illustrates the distribution of various modalities in diversity. VAE-AMDT is designed to transfer unimodal representations to a joint sentiment embedding space. As a result, we obtain discriminative sentiment multimodal representations and make it easier to predict sentiment intensity. "△" and "◯" indicate "non-negative" and "negative" respectively.

unimodal representations.) and the unified multimodal labels are not always adaptable to unimodal learning (*e.g.*, a unified multimodal label is *highly negative*, but text represents *neutral*), the distance difference between unimodal representations increases, and prevents systems from learning discriminative multimodal representations. The work [60] propose an adversarial encoder-decoder-classifier framework to reduce the modality gap by using adversarial training [61, 62], and the work [63] design a unimodal label auto-generation module to better learn unimodal representations for multimodal fusion. These two methods reduce the distance difference between unimodal representations via different approaches and aim to map various modalities in a joint embedding space so that the model can easily learn a common classifier. However, from the evaluation result, their efficacy is limited on the small and imbalanced sentiment dataset.

In this chapter, to obtain more discriminative multimodal representations that can further improve the performance of video-level sentiment analysis, as shown in Fig. 3.1, we propose a VAE-based adversarial multimodal domain transfer (VAE-AMDT) to better reduce the distance difference between unimodal representations and transfer various modalities to a joint embedding space, so that the model can easily learn discriminative multimodal representations and find an effective classifier over various modalities. Variational auto-encoder (VAE) is an auto-encoder whose training is regularised so that the distributions returned by its encoder are enforced to be close to a standard normal distribution [64, 65]. We perform it with visual, linguistic, and acoustic modalities respectively to make encoded latent representations follow a common distribution so that the modality gap can be reduced. Furthermore, motivated by [60], we introduce a discriminator trained with adversarial loss to classify encoded latent representations of target modality as true but others as false. As a result, we can better transfer encoded latent representations from various modalities to a joint embedding space as shown in Fig. 3.1. Then, we jointly train VAE-AMDT with a multi-attention module on this joint embedding space to learn more discriminative multimodal representations. The multi-attention module consists of self-attention, cross-attention, and triple-attention components, we employ it to highlight important sentimental representations over time and modality. Especially, we perform the cross-attention component under a "non-alignment" modality data setting to make our method can capture sequence-level interactions between modalities and have a much better multimodal fusion ability (*e.g.*, text → audio) [59]. We also perform self-attention to highlight important elements in each modality, and triple-attention to highlight important modality.

We conduct detailed experiments on the video-level sentiment analysis dataset MOSI [2] and MOSEI [3]. Our method improves the F1-score of the state-of-the-art method Self-MM [63] by **3.6%** on MOSI and **2.9%** on MOSEI datasets respectively. We also perform quantitative and qualitative analysis on the test set of both datasets, and the results suggest that VAE-AMDT is capable of reducing distance difference among unimodal representations, and fused multimodal representation is discriminative for improving the performance of video-level sentiment analysis.

3.2 Related work

Unimodal sentiment analysis

Sentiment analysis from people's facial expressions, voices, and speech texts has some impressive progress by employing deep learning techniques [56]. Convolutional neural networks (CNN) are employed to do facial expressions recognition (FER) [66, 67]; Recurrent neural networks (RNN) are employed to do speech emotion recognition (SER) [68, 69, 70, 71]; Language models (*e.g.*, BERT [72]) are finetuned to do textual sentiment analysis [73, 74, 75]; All these methods focus on learning effective latent representations from single modality. However single modality is not enough to provide comprehensive information to analyze people's complex sentiments. In contrast, our method focuses on how to fuse these unimodal latent representations to further improve the performance of sentiment analysis.

Multimodal fusion

Recent works on video-level sentiment analysis are increasing, and aim to gain more effective multimodal representations from various modalities. Several recent works [40, 57, 58] employ attention mechanisms to fuse multimodal representations through modeling interactions across various modalities. The work [3] proposes a dynamic fusion graph to do inter-multimodal fusion and the work [58] dynamically adjusts word representations using its aligned facial expressions and voice representations. However, these methods work with the forced alignment data setting and are limited to building sequence-level interactions between modalities. Our method works with a non-alignment data setting, so we can use cross-attention to build sequence-level optimal interactions cross modality.

To further improve the performance of multimodal fusion, recent works [60, 63] focus on how to reduce distance difference of unimodal representations since it is hard for systems to learn a common classifier from various modality domains as shown in Fig. 3.1. Motivated by adversarial training [76, 62], The work [60] introduce an adversarial encoder-decoder-classifier framework to transfer unimodal representations to a joint embedding space, and the work [63] designs a unimodal label auto-generation module to better learn unimodal representations so that the distance difference between modality can be reduced. However, their efficacy is limited on the small and imbalanced sentiment dataset. We perform adversarial training by using VAE-encoded unimodal representations to better reduce the distance difference of unimodal representations.

3.3 Problem Statement

In this chapter, we aim to predict people's sentiment intensity with a given video clip. The video clip includes multimodal signals: people's face image frames (I_v), audio (I_a), and speech text (I_t). We regard this task as a regression task, and our model takes I_v, I_a, and I_t as inputs and outputs one sentiment intensity $y \in R$. Here, R is in the range of $[-3, 3]$.

3.4 Modality data preprocessing

Given a video clip, we first drop out data that does not contain all of I_v, I_a, and I_t to ensure our model works properly, and then we process each unimodal signal following the below techniques to obtain their sequence features:

1. For the visual modality, we first use OpenFace [77] to extract I_v, and then we initialize visual sequence features $V \in \mathbb{R}^{T_v \times D_v}$ by encoding facial expression representations from I_v using a pretrained FER model [78]. Here, the FER model is pretrained on the VGG-Face dataset [79]. Given an extracted face image, we perform that pretrained FER model and use its prediction result as facial expression representations. The facial expression result is represented with an 8-dimensional vector. More details on Albanie's website [1].

2. For the linguistic modality, we initialize language sequence features $L \in \mathbb{R}^{T_l \times D_l}$ by extracting sentence embeddings of I_t using a pretraining language model RoBERTa [5].

3. For the acoustic modality, we initialize audio sequence features $A \in \mathbb{R}^{T_a \times D_a}$ by extracting log-mel filter banks from I_a [69].

In this chapter, to solve one problem of different video clip lengths, we do padding and truncation to adjust the length of V, L, and A respectively. We set T_v, T_l and T_a to 64, 100 and 128, and D_v, D_l and D_a to 8, 1024 and 128.

[1] https://www.robots.ox.ac.uk/~albanie/mcn-models.html

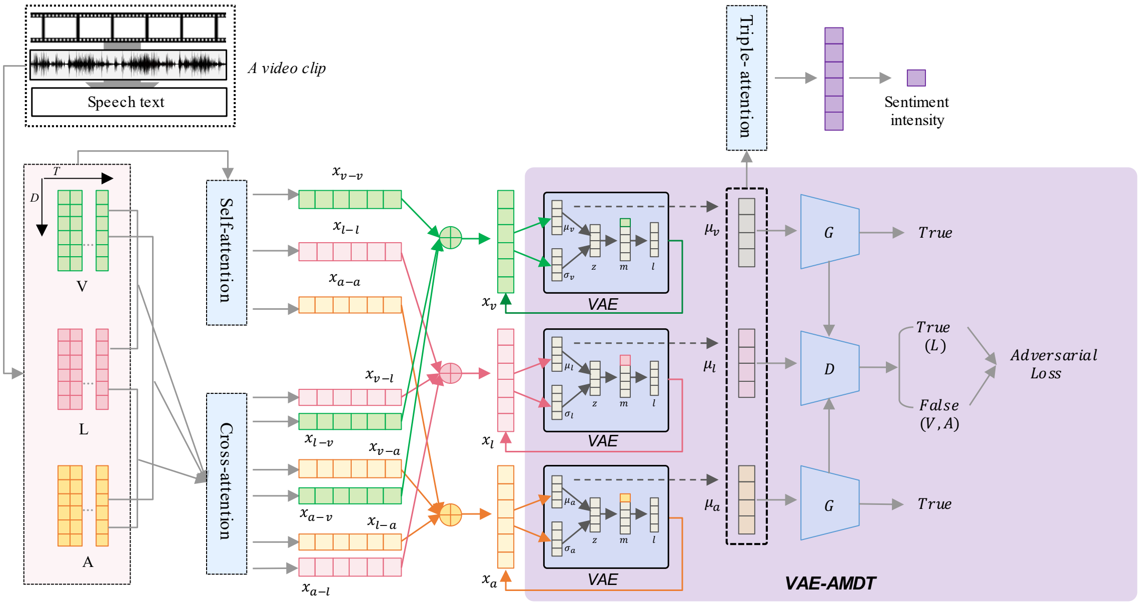


Figure 3.2: Overview of our method: we first perform self-attention (§3.5.1) and cross-attention (§3.5.1) using preprocessed sequence features V, L and A, and then we perform VAE-AMDT that consisting of three VAEs and two generators G and one discriminator D to reduce distance difference between unimodal representations (§3.5.2). Finally, we use the encoded unimodal representations as the input of triple-attention (§3.5.1) to output one sentiment intensity result. Here, unimodal representations x_v, x_l, and x_a indicate concatenations of the output of attention layers for each modality. μ_v, μ_l and μ_a are encoded unimodal representations with VAE-AMDT.

3.5 Methodology

In this section, we explain our method in detail. As shown in Fig. 3.2, our method includes VAE-AMDT and a multi-attention module that consists of self-attention, cross-attention, and triple-attention components. We jointly train VAE-AMDT and the multi-attention module to reduce the distance difference between unimodal representations and fuse multimodal representations to do sentiment intensity prediction.

3.5.1 Multi-attention module

self-attention The self-attention is designed to highlight key sequence elements [21, 72], and performed by taking V, A and L as inputs and output self-attention vector

$x_{(v \to v)}$, $x_{(l \to l)}$ and $x_{(a \to a)}$, as follows:

$$X_{(m)} = f_m(X) \tag{3.1}$$

$$x_{(m \to m)} = f_s\left(\frac{\sum_{t=1}^{T_m} \alpha_{(m \to m)} \cdot X_{(m)}}{T_m}\right), \alpha_{(m \to m)} = \text{softmax}(X_{(m)} \cdot X_{(m)}^T) \tag{3.2}$$

where $f_m : \mathbb{R}^{T_m \times D_m} \to \mathbb{R}^{T_m \times D}$ is a linear transformation. We perform f_m with $X \in \{V, L, A\}$ to output $X_{(m)}, m \in \{v, a, l\}$ and they have a same dimension D. We then calculate attention weight $\alpha_{(m \to m)}$ and get self-attention vector $x_{(m \to m)}$ via a 2-layer MLP $f_s : \mathbb{R}^D \to \mathbb{R}^D$.

Cross-attention We perform cross-attention between any two modalities to highlight correlated sequence elements over modality. For example, corresponding to one speech text “I enjoyed the party today.”, the word “enjoy” should attend to the enjoyable facial expressions, and its cross-attention weight $\alpha_{(m1 \to m2)}$ should be learned with a high score. We use $m1$ and $m2$ to indicate different modalities. We perform cross-attention in two attentional directions to get cross-attention vector $x_{(m1 \to m2)}$ and $x_{(m2 \to m1)}$, as follows:

$$x_{(m1 \to m2)} = f_s\left(\frac{\sum_{t=1}^{T_{m2}} \alpha_{(m1 \to m2)} \cdot X_{(m2)}}{T_{m2}}\right), \alpha_{(m1 \to m2)} = \text{softmax}(X_{(m1)} \cdot X_{(m2)}^T) \tag{3.3}$$

As shown in Fig. 3.2, we concatenate self-attention and cross-attention vectors for each modality to get unimodal representations x_m, as follows:

$$x_v = [x_{(v \to v)} || x_{(l \to v)} || x_{(a \to v)}] \tag{3.4}$$

$$x_l = [x_{(l \to l)} || x_{(v \to l)} || x_{(a \to l)}] \tag{3.5}$$

$$x_a = [x_{(a \to a)} || x_{(l \to a)} || x_{(v \to a)}] \tag{3.6}$$

where “$||$” denotes concatenation operation. We take x_v, x_l and x_a as inputs of VAE-AMDT (§3.5.2).

Triple-attention We fuse VAE-AMDT encoded unimodal representations μ_v, μ_l, and μ_a by using triple-attention so that the important unimodal representations can be highlighted. We stack μ_v, μ_l, and μ_a in a list and then perform Eqs. 3.1 and 3.2 to get a multimodal representation vector x. Finally, we perform linear regression for sentiment intensity prediction by employing mean squared error (MSE) loss function $\mathcal{L}_m$, as follows:

$$\mathcal{L}_m(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} |f_r(x_i) - y_i|^2 \tag{3.7}$$

where $f_r : \mathbb{R}^D \to \mathbb{R}^1$ is a linear transformation, used to output one sentiment intensity result. $\mathbf{x}$ and $\mathbf{y}$ represent a batch size n of the multimodal representation and ground truth label.

3.5.2 VAE-AMDT

VAE-AMDT is composed of three VAEs and two generators G and one discriminator D (Fig. 3.2). We jointly train it with the multi-attention module to transfer x_v, x_l, and x_a to a joint embedding space and use its output μ_v, μ_l and μ_a to predict sentiment intensity (§3.5.1). We show how to learn VAEs and how G and D worked in the adversarial training process as follows:

Variational auto-encoder (VAE) The Kullback-Leibler Divergence (KLD) term of VEA allows us to regularize the encoder to produce a latent vector z that follows a standard normal distribution [64, 65]. As a result, we have each mean layer $\mu_{(\mathbf{m})}$ that follows a similar distribution [64]. To further include modality type information in the encoder, we define a one-hot vector c_m representing different modality types and concatenate them with z to generate a modality conditional vector m for each modality. Here, (1,0,0), (0,1,0), and (0,0,1) represent three types of modalities. We define an MLP layer l as a decoder P_θ, and maximize the loss function $\mathcal{L}_{vae}$ to learn VAEs together as follows [65]:

$$\mathcal{L}_{vae}(\theta, \phi) = \sum_{r=1}^{R}\sum_{n=1}^{N}\{-\beta KL(Q_\phi(\mathbf{z}|\mathbf{x}_n^r)||P_\theta(\mathbf{z})) + \mathbb{E}_{Q_\phi(\mathbf{z}|\mathbf{x}_n^r)}\left[\log P_\theta(\mathbf{x}_n^r|\mathbf{z}, \mathbf{m})\right]\} \tag{3.8}$$

where ϕ and θ denote the parameters of the encoder and decoder respectively. R denotes the number of modalities and N denotes the data size. We set β to 0.5. This is a trade-off coefficient that allows the model to prioritize one term over the other. KL represents the KLD term, used to constrain the variational posterior $Q_\phi(\mathbf{z}|\mathbf{x})$ close to the prior $P_\theta(\mathbf{z})$. The second term on the right-hand side of Eq. 3.8 indicates the values of the expected log-likelihood generated by the decoder P_θ. To maximize it to enforce $\mathbf{z}$ return to the original data space with the constraint $\mathbf{m}$. Here, $\mathbf{m}$ represents a batch of modality conditional vectors. When KL is minimized, the encoder Q_ϕ is also constrained by $\mathbf{m}$. As a result, the type of modality can affect the encoder optimization and make the encoder represent the type of modality as well.

Adversarial training We take VAE encoded unimodal representations μ_v, μ_l, and μ_a as the input. To further reduce the distance between any two unimodal representations, we introduce two G to generate fake linguistic modal representations from visual and acoustic modality and then design a D to discriminate the real linguistic modal representation from generated fake representations by employing an adversarial loss $\mathcal{L}_{at}$. In addition, we perform binary classification for the generator by using binary cross entropy loss (BCELoss). We jointly train two G and one D to as follows:

$$\begin{gathered}\mathcal{L}_m^G = \arg\min_{E_m} V(E_m), \\ V(E_m) = \mathbb{E}_{\mu_{\mathbf{m}}\sim Q_{\psi_m}(\mu_{\mathbf{m}})}\left[\log E(\mu_m)\right] + \mathbb{E}_{\mu_m\sim Q_{\psi_m}(\mu_m)}\left[\log\left(1 - E_m(\mu_m)))\right]\end{gathered} \tag{3.9}$$

where E_m indicates generator of modality $m \in \{v, a\}$.

$$\begin{aligned}\mathcal{L}^D &= \arg\min_{E_v,E_a}\max_{D} V(E_v, E_a, D), \\ V(E_v, E_a, D) &= \mathbb{E}_{\mu_{\mathbf{l}}\sim Q_{\psi_l}(\mu_{\mathbf{l}})}\left[\log D(\mu_{\mathbf{l}})\right] + \mathbb{E}_{\mu_{(v)}\sim Q_{\psi_v}(\mu_{(v)})}\left[\log\left(1 - D\left(E_v(\mu_{(v)})\right)\right)\right] \\ &\quad + \mathbb{E}_{\mu_{(a)}\sim Q_{\psi_a}(\mu_{(a)})}\left[\log\left(1 - D\left(E_a(\mu_{(a)})\right)\right)\right]\end{aligned} \tag{3.10}$$

Consequently, we have $\mathcal{L}_{at}$ for adversarial training.

$$\mathcal{L}_{at} = \mathcal{L}_v^G + \mathcal{L}_a^G + \mathcal{L}^D \tag{3.11}$$

3.5.3 Learning

We finally have a joint loss $\mathcal{L}$ for training the multi-attention module and VAE-AMDT, as follows:

$$\mathcal{L} = \alpha\mathcal{L}_m + \beta\mathcal{L}_{ave} + \gamma\mathcal{L}_{at} \tag{3.12}$$

where α, β, and γ are hyperparameters, which are used to indicate the importance of each loss value. We empirically set them as 1.

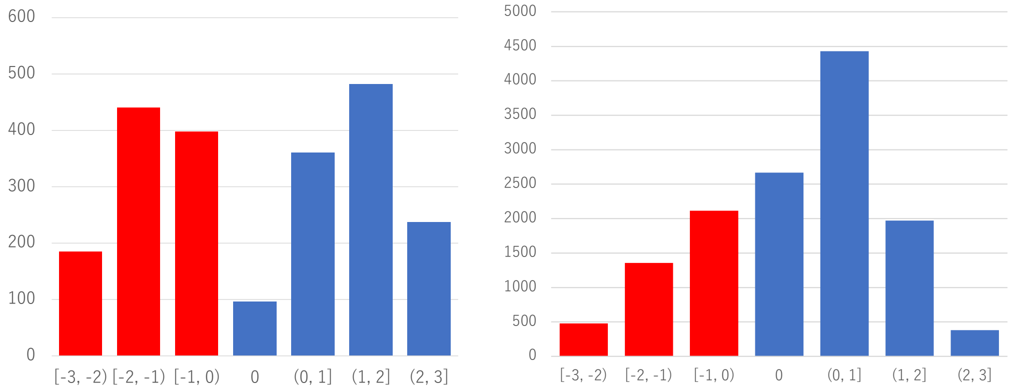
3.6 Experiment

3.6.1 Dataset

We evaluate our method by using video-level sentiment analysis datasets MOSI [2] and MOSEI [3]. Both datasets are collected from online video: MOSI contains 2,199 opinion video clips and MOSEI contains more than 65 hours of video from more than 1000 speakers and 250 topics. To ensure our method behaves correctly, we drop out data that does not contain all of the modalities. Tab. 3.1 shows the number of data in both datasets in detail. The MOSEI dataset is over 6x larger than the MOSI dataset. Both datasets are annotated in the range of the [-3,3] likert scale, *i.e.*, [-3: highly negative, -2: negative, -1: weakly negative, 0: neutral, +1: weakly positive, +2: positive, +3: highly positive]. From the data distribution over annotations in Fig. 3.3, we have very imbalanced data annotations for both datasets. In particular, there is over 65% of MOSEI dataset is annotated in the range of [-1, 1].

Dataset	Train	Validation	Test	Total
MOSI	1,257	229	686	2,172
MOSEI	9,473	1,206	2,710	13,389

Table 3.1: The size of datasets.



(a) MOSI

(b) MOSEI

Figure 3.3: Annotation distributions on (a) MOSI and (b) MOSEI. We show “negative” classes in red color and “non-negative” classes in blue color.

3.6.2 Metric

We use the mean absolute error (MAE), accuracy (A^2), and weight $F1$ score as evaluation metrics. A^2 is a binary accuracy metric, the prediction result $y < 0$ are belonged to "Negative" class and $y \geq 0$ are belonged to "Non-negative" class (Fig. 3.3). Furthermore, due to the small and imbalanced dataset, we also use the precision-recall curve to show the model's performance at various threshold settings.

3.6.3 Full model hyperparameters

We show full hyperparameters of our model on MOSI and MOSEI datasets in Tab. 3.2. We use AdamW [80] as our optimizer, with ϵ=1e-8. We use cosine annealing schedular [81] to adjust the learning rate (1e-8). We also show the feature size of each attention component in our multi-attention module (Fig. 3.2) in detail. Our hidden layer size (f_m) is different from datasets, so we have different hyperparameters for training their best performance (Tab.3.2: "Training").

3.6.4 Performance

As shown in Tab. 3.3, under the same modality alignment setting (non-alignment), our method achieves a much lower MAE result than Self-MM(+) by over ***0.16*** (MOSI) and ***0.05*** (MOSEI). Especially, a low MAE indicates that our method is superior to the sentiment regression problem. Compared to Self-MM(+), we also note that our method improves MAE results better with MOSI than with MOSEI. This suggests that VAE-AMDT is much more effective for relatively small datasets (Tab. 3.1). Here, Self-MM(+) is trained by using the same preprocessed data in our method (§3.4). Especially, we use the same pretrained RoBERTa model to encode speech text for a fair comparison between Self-MM(+) and our approach. We take out the data that lacks some modalities so that we can fairly compare their performance in terms of the modal fusion capability. To fairly confirm the binary classification ability of models trained with imbalanced annotations (Fig. 3.3), in addition to the accuracy (A^2) comparisons, we also show precision-recall curve for both MOSI and MOSEI in Fig. 3.4. The results suggest that our method is superior to Self-MM(+). Even though Self-MM(+)'s A^2 result (84.6%) is higher than our method (82.8%), when both precision and recall scores are over 80% as shown in the

[2] https://github.com/TadasBaltrusaitis/OpenFace

[3] https://huggingface.co/docs/transformers/v4.17.0/en/model_doc/roberta

		MOSI	MOSEI
Audio	Sample rate	44.1KHz	
	FFT hop length	0.02s	
	FFT window size	0.01s	
	Mel bins	128	
	Sequence length	128	
Image	Frame rate	8fps	
	Face detection	OpenFace [2]	
	Face frame size	128*128	
	Facial expressions feature (dim)	8	
	Sequence length	64	
Text	Tokenization	Roberta Tokenization [3]	
	Embeddings(dim)	1024	
	Sequence length	100	
f_m	Feature size (input)	V:(B, 64, 8); L:(B, 100, 1024); A:(B, 128, 128)	V:(B, 64, 8); L:(B, 100, 1024); A:(B, 128, 128)
	Feature size (output)	V:(B, 64, 128); L:(B, 100, 128); A:(B, 128, 128)	V:(B, 64, 64); L:(B, 100, 64); A:(B, 128, 64)
Self	Feature size (input)	V:(B, 64, 128); L:(B, 100, 128); A:(B, 128, 128)	V:(B, 64, 64); L:(B, 100, 64); A:(B, 128, 64)
	Feature size (output)	V:(B,128); L:(B, 128); A:(B, 128)	V:(B,64); L:(B, 64); A:(B, 64)
Cross	$[v \rightleftharpoons l]$ Feature size (input)	V:(B, 64, 128); L:(B, 100, 128)	V:(B, 64, 64); L:(B, 100, 64)
	$[v \rightleftharpoons l]$ Feature size (output)	$[v \to l]$ and $[l \to v]$:(B, 128)	$[v \to l]$ and $[l \to v]$:(B, 64)
	$[v \rightleftharpoons a]$ Feature size (input)	V:(B, 64, 128); A:(B, 128, 128)	V:(B, 64, 64); A:(B, 128, 64)
	$[v \rightleftharpoons a]$ Feature size (output)	$[v \to a]$ and $[a \to v]$:(B, 128)	$[v \to a]$ and $[a \to v]$:(B, 64)
	$[a \rightleftharpoons l]$ Feature size (input)	A:(B, 128, 128); L:(B, 100, 128)	A:(B, 128, 64); L:(B, 100, 64)
	$[a \rightleftharpoons l]$ Feature size (ouput)	$[a \to l]$ and $[l \to a]$:(B, 128)	$[a \to l]$ and $[l \to a]$:(B, 64)
Joint	Feature size (input)	x_v, x_l and x_a:(B, 384)	x_v, x_l and x_a:(B, 192)
Triple	Feature size (input)	(B,3,32) $[\mu_v, \mu_l, \mu_a]$	
	Feature size (output)	(B,32) $[x]$	
Optimizer	Peak learning rate	1e-4	
	Weight decay	0	
	AdamW β	0.9	
	AdamW ϵ	1e-8	
	Schedular	CosineAnnealingLR	
Training	Loss function	Mean Squared Error (MSE)	
	GPU	GTX 1080 Ti	
	Batch size	4	20
	Training epochs	200	80
	Parameters	3.3M	1.7M
	Training time	1h13m	46m
	Inference time	0.000738	0.000125
	Training time (Self-MM)	3h29m	-
	Inference time (Self-MM)	0.001131	-

Table 3.2: Full hyperparameters for our model. "dim" indicates the number of dimensions. "Self", "Cross" and "Triple" indicate self-attention, cross-attention, and triple-attention components respectively.

precision-recall curve graph (Fig. 3.4b), our method is still better than Self-MM(+). We also show the result of our method trained in a 10-fold cross-validation strategy (CV), which is a bit worse due to the small and imbalanced dataset, but it is still better than Self-MM(+) except A^2 for MOSEI. This result also suggests that our method is not overfitting to the training set.

We additionally compare the number of parameters of Self-MM and our method. Self-MM finetunes the pretrained BERT model [72], so it needs to reuse and update BERT's parameters, and the training parameters exceed ***100M***. This is ***33X*** larger than our method ***(3.3M)***. Since we utilize the pretrained RoBERTa [5] to embed speech text during preprocessing (§ 3.4), it is not essential to update massive pretrained parameters. As a result, we can not only train our method in a short time ($\frac{1}{3}$th of Self-MM) as shown in Tab. 3.2 but also achieves a model that is ***1.5X*** faster than Self-MM for inference.

Model	**MOSI**			**MOSEI**			**Modality alignment**
	MAE	A^2	F1	MAE	A^2	F1	
Graph-MFN	0.965	77.4	77.3	-	76.0	76.0	Yes
RAVEN	0.915	78.0	76.6	0.614	79.1	79.5	Yes
ARGF	-	81.3	81.5	-	-	-	Yes
MulT	0.861	81.5	80.6	0.580	-	-	Yes
MISA (*)	0.804	80.8	80.8	0.568	82.6	82.7	Yes
MAG-BERT (*)	0.731	82.5	82.6	0.539	83.8	83.7	Yes
Self-MM (*)	0.713	84.0	84.4	0.530	82.8	82.5	No
Self-MM (+)	0.885	80.6	80.6	0.579	**84.6**	84.6	No
VAE-AMDT	**0.716**	**84.3**	**84.2**	**0.526**	82.8	**87.5**	No
VAE-AMDT (CV)	0.745	82.2	82.2	0.529	81.6	86.2	No
Human	0.710	85.7	87.5	-	-	-	No

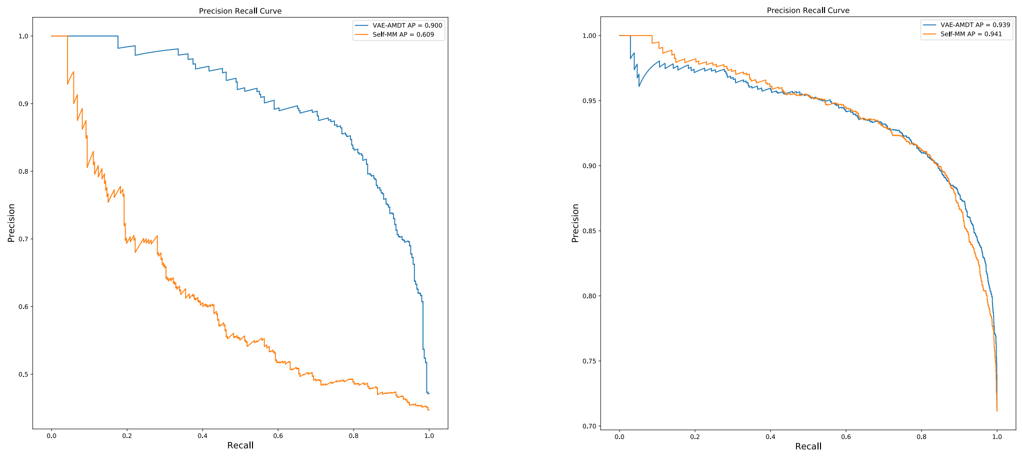
Table 3.3: Comparison of *VAE-AMDT* and state-of-the-art results in both MOSI and MOSEI. *VAE-AMDT* outperforms state-of-the-art Self-MM (*MAE/F1*) by over ***0.16/3.6*** point (MOSI) and ***0.05/2.9*** point (MOSEI). Here, the lower the MAE, the better the performance. (*) indicates that the results are referenced from the Self-MM paper. (+) indicates that Self-MM is trained by using the same preprocessed data in our method; (CV) indicates the result of the 10-fold cross validation.

3.6.5 Effect of VAE-AMDT

We first show the comparison results of our method built (w/o and w/) *VAE-AMDT* in Tab. 3.4. The results suggest that our proposed *VAE-AMDT* is effective for improving the performance of the model only built by employing the multi-attention module (§ 3.5.1). We further study the effect of *VAE-AMDT* through quantitative and qualitative analysis.

Model	**MOSI**			**MOSEI**			**Modality alignment**
	MAE	A^2	F1	MAE	A^2	F1	
w/o VAE-AMDT	0.808	80.3	80.6	0.603	81.8	85.8	No
w/ VAE-AMDT	**0.716**	**84.3**	**84.2**	**0.526**	**82.8**	**87.5**	No

Table 3.4: Comparison results of the model trained w/o and w/ VAE-AMDT. The model trained with VAE-AMDT further improves F1 score of (w/o VAE-AMDT) by 3.6% (MOSI) and 1.7% (MOSEI).



(a) MOSI (b) MOSEI

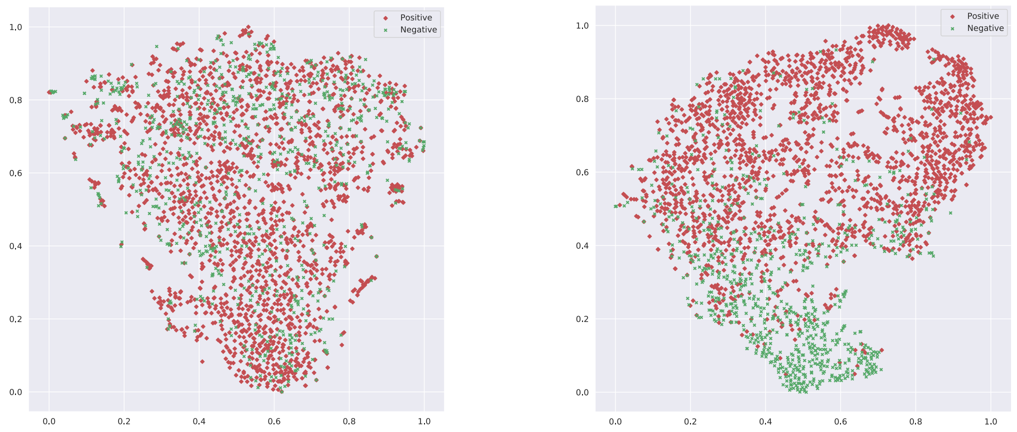
Figure 3.4: The precision-recall curve is created by using VAE-AMDT's test prediction results on both datasets. The curve indicates that VAE-AMDT outperforms Self-MM when both precision and recall scores exceed 0.8. Here, a better model should perform better for both metrics.

Maximum mean discrepancy score(MMD) We do quantitative analysis by analyzing the maximum mean discrepancy (MMD) on both MOSI and MOSEI test sets. The MMD is a kernel-based approach that is used to measure the distance between two probability distributions [82]. We use encoded unimodal representations μ_v, μ_l, and μ_a to calculate the MMD score between any two modalities and show their results in Tab. 3.5. Our proposed *VAE-AMDT* not only can balance the distance difference between any modality pairs (*e.g.*, v→l, a→l and v→a), but also reduce their average distance difference in total and prove the efficacy of *VAE-AMDT*.

Method	**MOSI**				**MOSEI**			
	$v \to l$	$a \to l$	$v \to a$	*Average*	$v \to l$	$a \to l$	$v \to a$	*Average*
w/ AMDT	**0.51**	**0.17**	1.44	0.71	0.98	0.92	0.45	0.79
w/ VAE-AMDT	0.68	0.53	**0.33**	**0.51**	**0.28**	**0.30**	**0.25**	**0.28**

Table 3.5: MMD results show that not only can the model (w/ VAE-AMDT) balance the distance between any two modalities, but the average result is lower than the model (w/ AMDT).

Visualization To further explain the efficacy of *VAE-AMDT*, we perform qualitative analysis by visualizing the encoded unimodal representations using t-SNE and show the result on the MOSEI test set in Fig. 3.5. We concatenate encoded unimodal representations μ_v, μ_l, and μ_a and use t-SNE to map them into a joint embedding space. By applying *VAE-AMDT* (Fig. 3.5b), the dots indicating "negative" and "non-negative" classes tend to split into two clusters and prove that *VAE-AMDT* is capable of obtaining discriminative multimodal representations.



(a) w/ AMDT

(b) w/ VAE-AMDT

Figure 3.5: Visualization result on MOSEI. The green color indicates the "negative" class and the red color indicates the "positive" (including "neutral") class. The model (w/ VAE-AMDT) classifies both classes by discriminative representations.

3.6.6 Ablation study

To prove the efficacy of all components in our method, we study the Multi-attention module and Modality respectively. Here, we discuss all comparison results based on the MAE metric. We consider that the MAE metric should be more reliable than the A2 and F1 metric on regression learning, especially for small and imbalanced datasets.

Multi-attention module To confirm the effect of all components of the multi-attention module, we show the comparison results of the model trained by employing different attention components in Tab. 3.6. For the model employing triple-attention w/o *VAE-AMDT*, we use unimodal representations x_v, x_l and x_a instead of μ_v, μ_l and μ_a. The result suggests that (self, cross, triple)-attention improves performance when used together. Especially, the MAE result is improved much after adding triple-attention and suggests its efficacy in highlighting the important modality.

Attention type	Sentiment intensity		
	MAE	A^2	$F1$
self-attention	0.688	80.2	85.5
(self, cross)-attention	0.683	81.2	**86.3**
(self, cross, triple)-attention	**0.603**	**81.8**	85.8

Table 3.6: Ablation study of the multi-attention module on MOSEI dataset. All models are trained w/o VAE-ADMT, (self, cross, triple)-attention shows the lowest MAE score compared to others.

Modality To ensure that increasing the number of modalities can improve performance, we compare the models that are trained given various modalities as the input and show the results in Tab. 3.7. It is clear that adding modality improves performance. However, we note that speech text performs better than other modalities (*e.g.*, image, audio). We believe that the language encoder (RoBERTa model [5]) we used is more powerful than encoders used for image and audio.

Modality	MOSI			MOSEI		
	MAE	A^2	$F1$	MAE	A^2	$F1$
Image	1.467	43.4	57.3	0.854	70.0	82.4
Audio	1.498	46.9	54.7	0.867	70.0	83.4
Text	0.990	83.8	75.7	0.698	78.9	84.3
Image, Audio	1.473	52.0	56.2	0.837	67.9	78.6
Audio,Text	0.875	80.0	69.0	0.646	82.8	**87.9**
Image,Text	1.140	74.9	68.1	0.593	82.7	87.8
Image,Text,Audio	**0.716**	**84.3**	**84.2**	**0.526**	**82.8**	87.5

Table 3.7: Effect of modality. The test results show that adding modality improves performance.

3.6.7 Case study

We show some data samples from the MOSEI test set in Tab. 3.8. The predicted sentiment intensity by our method is close to ground truth. Although we select samples randomly the result suggests that our method performs stable with these data. Furthermore, we note that some predicted score is more reasonable than ground truth. For example, the sample (ID:5) is predicted to be 0.52, which is lower than ground truth. However, we note that the speech text represents negative sentiment. These results not only prove that our method is not overfitting to the training set but also suggest that it is robust to practical use.

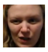
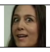
ID	Speech text	Face image	Sentiment intensity	
			Ground Truth	Prediction
1	This movie (umm) if you saw previews for it it looks kind of funny, but this movie wasn't very funny		-1.33	-1.69
2	On the other hand he's battling against his fellow atheists who deny that there are any objective moral values and duties		0.00	0.13
3	Millions of women, men, and children have better lives today thanks to the work that many of you have done for decades.		2.00	1.46
4	We will accomplish these goals by reviewing the law that pertains to mandatory child abuse reporting.		1.00	0.82
5	Get ready for this to be bad.“ That's not good, that's not what you want to do.		1.00	0.52
6	What I cared about was my finances and I felt like I was a pretty smart person and yet you know my financial life was not going where I wanted it to go.		-1.67	-1.27
7	I'm pleased that the United States is represented in Doha by Attorney General-Eric Holder and one of my key White House advisors, Mike Froman.		1.67	0.84

Table 3.8: Case study on MOSEI test set. The predicted sentiment intensity by our method is close to ground truth.

3.7 Conclusion

We proposed (*VAE-AMDT*) and jointly trained it with a multi-attention module to reduce the distance difference of various unimodal representations. As a result, we obtained discriminative multimodal representations to further improve the performance of video-level sentiment analysis. Our method balanced the distance difference between any modality pairs and reduced their average distance in total. We finally improve the F1-score of the state-of-the-art Self-MM by **3.6%** on MOSI and **2.9%** on MOSEI datasets, and prove the efficacy of our method in obtaining discriminative multimodal representations.

Chapter 4

VQA-GNN: Reasoning with Multimodal Knowledge via Graph Neural Networks for Visual Question Answering

In the last chapter, we introduced an approach called VAE-AMDT to enforce multimodal representations following a common regular distribution. As a result, our approach improved multimodal fusion and obtained discriminative multimodal representations. In this chapter, we propose a bidirectional fusion approach to enable systems performing concept-level reasoning by unifying unstructured (*e.g.*, the context in question and answer; "QA context") and structured (*e.g.*, knowledge graph for the QA context and scene; "concept graph") multimodal knowledge.

4.1 Introduction

The visual question answering (VQA) task aims to provide answers to questions about a visual scene. It is crucial in many real-world tasks including scene understanding, autonomous vehicles, search engines, and recommendation systems [12, 13, 14, 15]. To solve VQA, systems need to perform concept-level reasoning by unifying unstructured (*e.g.*, the context in question and answer; "QA context") and structured (*e.g.*, knowledge graph for the QA context and scene; "concept graph") multimodal knowledge.

Most of the high-performing VQA methods [24, 25, 26, 83, 84, 27, 28] pretrain a multimodal transformer model on a large-scale dataset to obtain unstructured multimodal knowledge from image and language contexts, and then finetune the pretrained model to reason on downstream tasks (*e.g.*, visual commonsense reasoning (VCR) task [85]). Existing methods (*e.g.*, SGEITL [86]) also incorporate structured knowledge into these transformer-based models by including a scene graph in the input of a pretrained multimodal transformer model. More recent methods [87, 88] further combine the scene graph and the concept graph by inter-connecting corresponding visual nodes and concept nodes through graph neural networks (GNNs) and then incorporate the unstructured QA context representation to perform question answering. However, these methods only perform late fusion or unidirectional fusion from unstructured knowledge to structured knowledge and do not train the model to mutually aggregate information from both sides. This can limit their potential to perform joint reasoning over the heterogeneous modalities of knowledge. As unstructured knowledge and structured knowledge have complementary benefits—pretrained unstructured representations capture broader knowledge and structured representations offer scaffolds for reasoning—[89], this motivates the development of models that deeply fuse the two modalities of knowledge for visual question answering.

We propose VQA-GNN (Fig. 4.1), a new visual question answering model performing bidirectional fusion between unstructured and structured multimodal knowledge to obtain a unified, more expressive knowledge representation. VQA-GNN extracts a scene graph from the given input image using an off-the-shelf scene graph generator [90] and then retrieves a relevant concept graph for the input image and QA context from a general knowledge graph like ConceptNet [91], obtaining a structured representation of the scene. Simultaneously, to obtain an unstructured knowledge representation for the scene, (1) we use pretrained RoBERTa [5] to encode the context in question and answer ("QA-context") as *QA-context node*, and (2) we retrieve relevant visual regions from a general scene graph VisualGenome [92] and take their mean pooled representation as a *QA-concept node*, which we connect to the scene graph. We then connect the scene graph and the concept graph through *QA-context node* to build a multimodal semantic graph.

To achieve bidirectional fusion across the multimodal semantic graph, we introduce a new multimodal GNN technique that performs inter-modal message passing. The multimodal GNN consists of two modality-specialized GNN modules, one for each modality, which perform inter-message aggregation between the *QA-context*

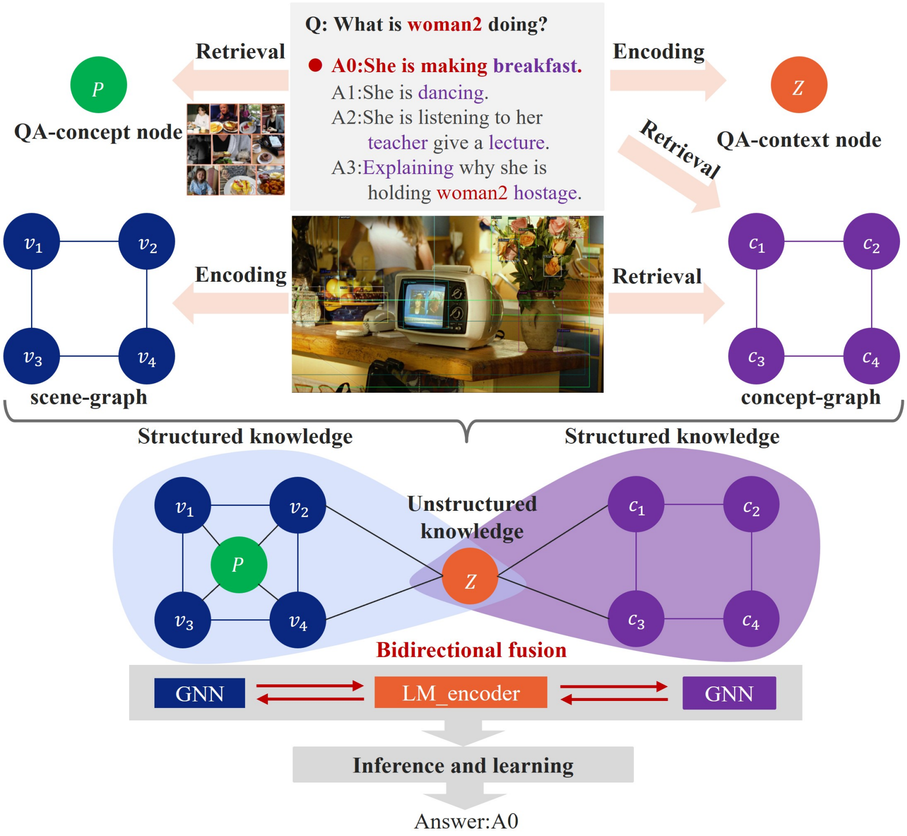


Figure 4.1: Overview of *VQA-GNN*. Given an image and QA sentence, we obtain unstructured knowledge (*e.g.*, QA-concept node p and QA-context node z) and structured knowledge (*e.g.*, scene-graph and concept-graph), and then unify them to perform bidirectional fusion for visual question answering.

node and nodes in structured graphs, aiming to reduce representational gaps between modalities. Meanwhile, by leveraging the robust transformer-based architecture of RoBERTa, we unfreeze and finetune the weights of the QA-context node to enable mutual information aggregation from modality-specialized GNN modules.

We evaluate VQA-GNN on two challenging VQA tasks, VCR [85] and GQA [93]. These tasks require systems to perform conceptual and compositional reasoning to answer diverse questions (*e.g.*, multiple-choice question answering and rationale selection in VCR; open-domain question answering in GQA). Our model outperforms strong baseline VQA methods [86, 94] by **3.2%** on VCR (Q-AR) and **4.6%** on GQA. Moreover, ablation studies show the efficacy of our two main techniques, bidirectional fusion and multimodal GNN message passing. On VCR, our multimodal GNN technique that reduces multimodal gaps outperforms existing works that use generic GNNs [87, 88] by **4.5%**. On GQA, bidirectional fusion outperforms a unidirectional fusion variant by **4%**. These results confirm the promise of VQA-GNN in unifying unstructured and structured multimodal knowledge for reasoning.

4.2 Problem Setup

This work focuses on multiple-choice and open-domain visual question answering, respectively. Each data point consists of an image c, and a natural language question q. For the multiple-choice setting, each question corresponds to a set of candidate answers $\mathcal{A}$, where only one candidate $a_{\text{correct}} \in \mathcal{A}$ is the correct answer to the question. Given a QA example $(c, q, \mathcal{A})$, we assume we have access to its relevant joint graph $\mathcal{G}^{(vcr)}$ and our goal is to identify the correct answer $a_{\text{correct}} \in \mathcal{A}$. For the open-domain setting, all questions correspond to a large set of common answer classes $\mathcal{B}$, where only one candidate $b_{\text{correct}} \in \mathcal{B}$ is the best answer to each question. Given a data example (c, q) with relevant scene graph $\mathcal{G}^{(gqa)}$, the goal is to identify $b_{\text{correct}} \in \mathcal{B}$.

4.3 Related Work

4.3.1 Multimodal transformer

VQA has emerged as one of the most popular topics in the computer vision community over the past few years [12, 13, 14, 15, 95, 96]. Existing methods for VQA [24, 25, 26, 27] employ the pretrain-and-finetune approach, where they train a multimodal transformer model on large-scale visual-language datasets, and then finetune the pretrained model on the downstream VQA datasets, *e.g.*, RESERVE-L model [28] is pretrained using 1 billion image-caption data including video frames, text, and audio. However, these methods only focus on obtaining unstructured multimodal representations by modeling implicit interactions over the visual and language domains. In contrast, our method introduces a multimodal GNN module to obtain unified knowledge representations from unstructured and structured multimodal knowledge based on explicit interactions over a well-structured multimodal semantic graph.

4.3.2 Structured knowledge-based VQA

Scene graph. Existing methods such as [83] introduce a scene graph prediction task to learn structured knowledge conditioned multimodal representations, and the work [86] proposes to incorporate extracted scene graph in multimodal transformer models. These works [97, 98, 94, 99] also exploit GNNs [100, 101, 102, 103] to incorporate unstructured QA-context knowledge into a structured scene graph for question answering. However, these methods only perform late fusion or unidirectional fusion from unstructured knowledge to structured knowledge. In contrast, our method performs bidirectional fusion to unify unstructured and structured knowledge.

Concept graph. Aiming to achieve concept-level reasoning beyond image-level recognition for visual understanding, existing works [104, 105, 106, 107, 108, 109, 87, 88, 110, 111, 112, 113] utilize knowledge graphs (KGs) to explore how to unify commonsense knowledge [89, 114, 115] about background concepts of the scene. The work [116] converts the image into captions and performs GPT-3 [22] in joint knowledge retrieval and reasoning. The work [111] encodes question-related knowledge from the retrieved knowledge facts to a knowledge-aware question representation, and then performs a question and knowledge-guided graph attention operation for answer reasoning. However, structured concept knowledge relevant to the QA context is not enough to represent the background scene. We build a concept graph to cover structured and unstructured concept knowledge relevant to the QA context as well as the background scene.

Scene graph & concept graph. To enrich structured knowledge, these works [107, 108, 110] utilize GNNs to learn graph representations of the scene graph and concept graph respectively, and then perform later fusion across the QA context, scene graph and concept graph for question reasoning. However, it is insufficient to capture the interactions across different modalities for concept-level reasoning. These works [87, 88] unify the scene graph and concept graph by interconnecting corresponding visual and concept nodes to capture their interactions. However, the representational gap between modalities adversely affects the performance of inter-modal message passing for capturing joint reasoning [117, 118]. Our method inter-connects the scene graph and concept graph via a QA context node and introduces a new multimodal GNN technique to mitigate representational gaps between modalities.

4.4 Methodology

As shown in Fig. 4.2, given an image and its related question with an answer choice, first we build a multimodal semantic graph to unify unstructured and structured multimodal knowledge into a joint graph (§4.4.1). Then we propose a multimodal GNN-based bidirectional fusion method that performs inter-modal message passing to obtain node representations enhanced with unstructured and structured multimodal knowledge (§4.4.2). Finally, we get the pooled representations of scene-graph and concept-graph and concatenate them with the representations from the QA-context node and QA-concept node for answer prediction (§4.4.3).

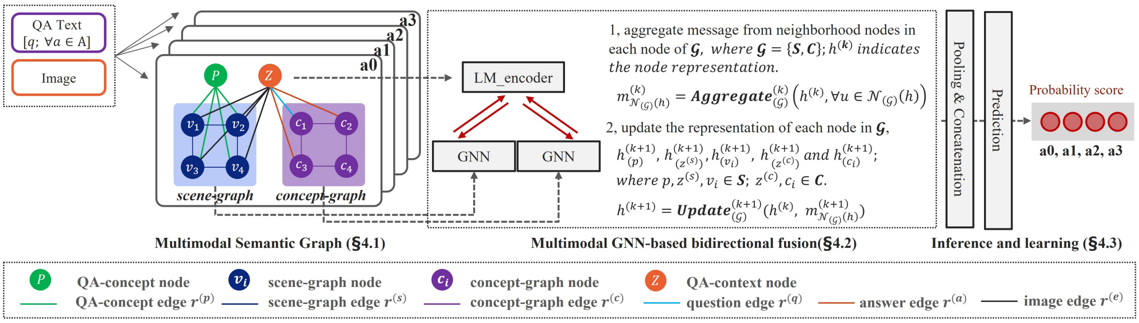

QA Text
$[q; \forall a \in A]$
Image
a3
a2
a1
a0
P
Z
v_1
v_2
v_3
v_4
c_1
c_2
c_3
c_4
scene-graph
concept-graph
LM_encoder
GNN
GNN
1, aggregate message from neighborhood nodes in each node of $\mathcal{G}$, where $\mathcal{G} = \{S, C\}$; $h^{(k)}$ indicates the node representation.
$m^{(k)}_{\mathcal{N}(\mathcal{G})(h)} = \textbf{Aggregate}^{(k)}_{(\mathcal{G})}\left(h^{(k)}, \forall u \in \mathcal{N}_{(\mathcal{G})}(h)\right)$
2, update the representation of each node in $\mathcal{G}$,
$h^{(k+1)}_{(p)}, h^{(k+1)}_{(z^{(s)})}, h^{(k+1)}_{(v_i)}, h^{(k+1)}_{(z^{(c)})}$ and $h^{(k+1)}_{(c_i)}$;
where $p, z^{(s)}, v_i \in S$; $z^{(c)}, c_i \in C$.
$h^{(k+1)} = \textbf{Update}^{(k+1)}_{(\mathcal{G})}(h^{(k)}, m^{(k+1)}_{\mathcal{N}(\mathcal{G})(h)})$
Pooling & Concatenation
Prediction
Probability score
a0, a1, a2, a3
Multimodal Semantic Graph (§4.1)
Multimodal GNN-based bidirectional fusion(§4.2)
Inference and learning (§4.3)
P QA-concept node
QA-concept edge $r^{(p)}$
v_i scene-graph node
scene-graph edge $r^{(s)}$
c_i concept-graph node
concept-graph edge $r^{(c)}$
Z QA-context node
question edge $r^{(q)}$
answer edge $r^{(a)}$
image edge $r^{(e)}$

Figure 4.2: Reasoning procedure of VQA-GNN. We first build a multimodal semantic graph for each given image-QA pair to unify unstructured (*e.g.*, "node p" and "node z") and structured (*e.g.*, "scene-graph" and "concept-graph") multimodal knowledge (§4.4.1). Then we perform inter-modal message passing with a multimodal GNN-based bidirectional fusion method (§4.4.2) to update the representations of node z, p, v_i and c_i for $k + 1$ iterations in two steps. Finally, we predict the answer with these updated various node representations (§4.4.3). Here, "S" and "C" indicate scene-graph and concept-graph respectively. "LM_encoder" indicates a language model used to finetune QA-context node representation, and "GNN" indicates a relation-graph neural network for iterative message passing.

4.4.1 Multimodal semantic graph

Scene-graph encoding. Given an image, we use a pretrained scene graph generator to extract a scene graph that consists of recall@20 of $(subject, predicate, object)$ triplets to represent structured image context [90], *e.g.*, $(car, behind, man)$. Then we apply a pretrained object detection model for embedding a set of scene graph nodes $\mathcal{V}^{(s)} = \{v_i\}_{i=1}^{N}$ (N indicates the maximum number of scene-graph nodes of "20") and represent $v_i^{(s)}$ with a 2048 dimensional visual feature vector [119]. We indicate the predicate edge types in the scene graph with a set of scene graph edges

$\mathcal{E}^{(s)} = \{r_i^{(s)}\}_{i=1}^{D}$ (D denotes the number of edge types) and represent $r_j^{(s)}$ with a D-dimensional one-hot vector.

QA-concept node retrieval. In addition to the local image context, with an assumption that the global image context of the correct choice aligns with the local image context, we employ a pretrained sentence-BERT model to calculate the similarity between each answer choice and all descriptions of the region image within the VisualGenome dataset [92]. This process allows us to extract relevant region images that capture the global image context associated with each choice [120]. We retrieve the top 10 results and utilize the same object detector to embed them. These embeddings are averaged to obtain a QA-concept node denoted as p. Subsequently, we introduce a QA-concept edge, denoted as $r^{(p)}$, which serves to fully connect node p with node v_i.

Concept-graph retrieval. We retrieve a concept graph from the image and ConceptNet KG, a general-domain knowledge graph [91]. Our process is illustrated in Fig. 4.3. In **Step 1**, we extract concept entities from both the image and the answer choices. Specifically, for the image, we consider the detected object names as potential contextual entities, while excluding general terms like "person" to streamline the reasoning process. For the answer choice, we ground phases if they are mentioned concepts in the ConceptNet KG, *e.g.*, "beverage" and "shop". In **step 2-1**, we use grounded phases to retrieve their 1-hop neighbor nodes from the ConceptNet KG. In **step 2-2**, since many concept nodes retrieved are semantically irrelevant to the answer choice, we use a word2vec model released by the spaCy library[1] to get relevance score between concept node candidates and answer choices, and prune irrelevance nodes when the relevance score is less than 0.6. As a result, given an answer choice, we can retrieve a relevance subgraph from ConceptNet KG based on the relevance score. In **step 3**, to better comprehend concept knowledge from the image as well, in addition to linking adjacent object entities in the ConceptNet KG domain, we also combine parsed local concept entities of the image with the retrieved subgraph. For instance, considering that ConceptNet encompasses various types of local concept entities, when a local concept entity (*e.g.*, "bottle") is found adjacent to a retrieved entity (*e.g.*, "beverage"), we build a new knowledge triple, *e.g.*, (bottle, allocation, beverage). Finally, we can construct a concept graph to depict the structured knowledge at the concept level. We obtain a collection of concept-graph nodes denoted as $\mathcal{V}^{(c)} = \{c_i\}_{i=1}^{N}$, where N represents the maximum number of concept-graph nodes

[1] https://spacy.io/

of 60. The concept entity c_i is represented using a 1024-dimensional text feature vector as the concept entity embedding in [121]. Additionally, we initialize a set of concept-graph edges denoted as $\mathcal{E}^{(c)} = \{r_i^{(c)}\}_{i=1}^{D}$, using D-dimensional one-hot vectors, where D is the number of edge types in concept-graph.

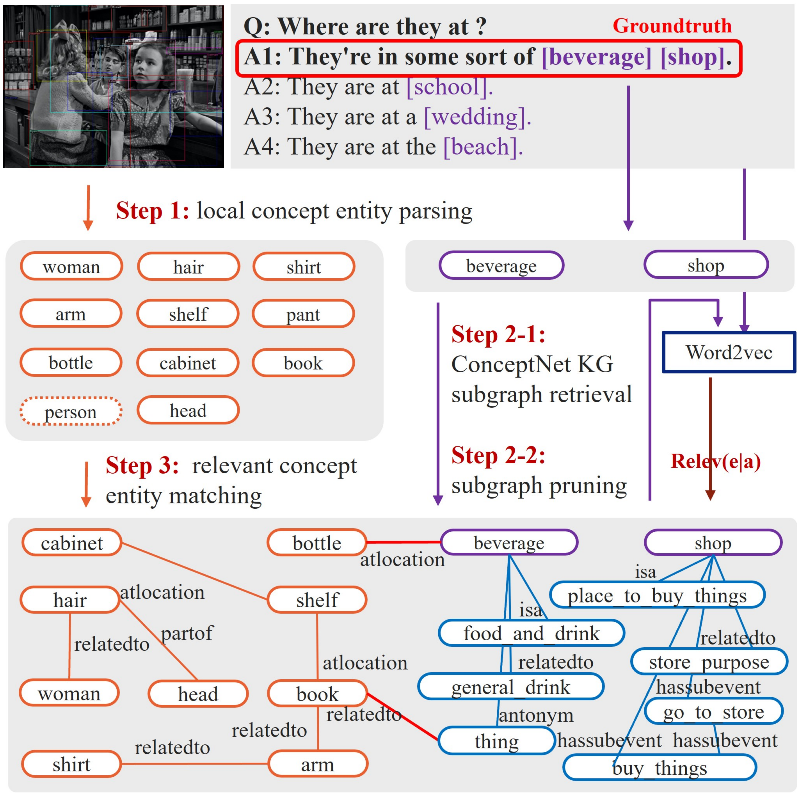


Figure 4.3: The process of concept-graph retrieval involves the calculation of similarity between concept-graph nodes and the answer context, denoted as $Relev(e|a)$.

QA-context node encoding. To construct a multimodal semantic graph, we introduce an unstructured QA-context node denoted as z to inter-connect the scene-graph and concept-graph using three additional relation types: the question edge $r^{(q)}$, the answer edge $r^{(a)}$, and the image edge $r^{(e)}$. The image edge $r^{(e)}$ fully links node z with $\mathcal{V}^{(s)}$, capturing the relationship between the QA context and relevant entities within the scene-graph. The question edge $r^{(q)}$ and answer edge $r^{(a)}$ link node z with the entities extracted from the question and the answer text, respectively, capturing the relationship between the QA context and the relevant entities within the concept-graph. As a result, we construct a multimodal semantic graph $\mathcal{G} = \{S, C\}$ to provide a joint reasoning space, which includes two sub-graphs of scene-graph S and concept-graph C, two super nodes of QA-concept node and QA-context node. Here, the QA-concept node is included in S and the QA-context is included in S and

C for performing inter-modal message passing in §4.4.2. Especially, the QA-context node z is assigned to not only learn unstructured discriminative representations by giving Q and A text pairs but also to incorporate structured multimodal knowledge from scene-graph and concept-graph for effective VQA. As the transformer-based method is powerful for multimodal representation learning [25, 24], we employ the RoBERTa LM [5] as the encoder of QA-context node z and finetune it with GNN modules to achieve bidirectional multimodal knowledge fusion (see Fig. 4.2).

4.4.2 Multimodal GNN-based bidirectional fusion

To improve inter-modal message passing by avoiding directly aggregating neighborhood nodes that may be initialized in different modality domains, we propose a multimodal GNN-based bidirectional fusion method built by two relation-graph neural networks for scene-graph and concept-graph respectively (see §4.4.1). The relation-graph neural network is built on the Graph Attention Networks (GAT) [101] by introducing multi-relation aware message for attention-based message aggregation process to better capture multiple relation information.

The details of the relation-graph neural network are as follows: we have four node types: $\mathcal{T} = \{\boldsymbol{Z}, \boldsymbol{P}, \boldsymbol{S}, \boldsymbol{C}\}$ in the multimodal semantic graph and they indicate QA-context node z, QA-concept node p, scene-graph node s, and concept-graph node c. As relation edge representation $\boldsymbol{r}_{i,j}$ should capture relationship from node i to node j and difference of node types represents a special relation between neighborhood nodes, we first obtain node type embedding u_i, u_j and then concatenate them with edge embedding e_{ij} to generate multi-relation embedding $\boldsymbol{r}_{ij}$ from i to j by

$$\boldsymbol{r}_{ij} = f_r([e_{ij}||u_i||u_j]) \tag{4.1}$$

where $u_i, u_j \in \{0,1\}^{|\mathcal{T}|}$ are one-hot vectors indicating the node types of i and j, $e_{ij} \in \{0,1\}^{|\mathcal{R}|}$ is a one-hot vector indicating relation type of edge (i,j). $||$ is the concatenation operation, and $f_r : \mathbb{R}^{|\mathcal{R}|+2|\mathcal{T}|} \rightarrow \mathbb{R}^{\mathcal{D}}$ is a 2-layer MLP. Based on multi-relation embedding $\boldsymbol{r}_{ij}$, the multi-relation aware message $\boldsymbol{m}_{ij}$ from i to j is computed by

$$\boldsymbol{m}_{ij} = f_m([\boldsymbol{h}_i^{(k+1)}||\boldsymbol{r}_{ij}]) \tag{4.2}$$

where $f_m : \mathbb{R}^{2\mathcal{D}} \rightarrow \mathbb{R}^{\mathcal{D}}$ is a linear transformation. $h_i^{(k+1)}$ is the node representation of each node i in the graph. We then recursively updated it $k+1$ times by

$$\boldsymbol{h}_i^{(k+1)} = f_h\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}\boldsymbol{m}_{ij}\right) + \boldsymbol{h}_i^{(k)} \tag{4.3}$$

where $f_h : \mathbb{R}^{\mathcal{D}} \to \mathbb{R}^{\mathcal{D}}$ is 2-layer MLP with batch normalization [122]. $\mathcal{N}_i$ indicates the neighborhood of node i, α_{ij} is an attention weight to emphasize important messages passed from $\mathcal{N}_i$ to node i. We obtain $\boldsymbol{q}_i, \boldsymbol{k}_j$ by

$$\boldsymbol{q}_i = f_q(\boldsymbol{h}_i^{(k+1)}), \boldsymbol{k}_j = f_k([\boldsymbol{h}_j^{(k+1)}||\boldsymbol{r}_{ij}]) \tag{4.4}$$

where $f_q : \mathbb{R}^{\mathcal{D}} \to \mathbb{R}^{\mathcal{D}}$ and $f_k : \mathbb{R}^{2\mathcal{D}} \to \mathbb{R}^{\mathcal{D}}$ are linear transformations. α_{ij} is computed using the softmax function by

$$\gamma_{ij} = \frac{\boldsymbol{q}_i^T\boldsymbol{k}_j}{\sqrt{D}}, \tag{4.5}$$

$$\alpha_{ij} = \text{softmax}_j(\gamma_{ij}) = \frac{\exp(\gamma_{ij})}{\sum_{j'\in\mathcal{N}_i}\exp(\gamma_{ij'})} \tag{4.6}$$

By referring to Eq. 4.3, we perform message passing to update node representations in each graph in parallel by aggregating multi-relation aware messages from neighborhood nodes in each node. As a result, we obtain structured graph node representations $h_{(v_i)}^{(k+1)}$ and $h_{(c_i)}^{(k+1)}$, unstructured node representations $h_{(p)}^{(k+1)}$ and $h_{(z)}^{(k+1)}$. For node z, we update it with scene-graph and concept-graph respectively, and concatenated by

$$\boldsymbol{h}_{(z)}^{(k+1)} = f_z([\boldsymbol{h}_{(z^{(s)})}^{(k+1)}||\boldsymbol{h}_{(z^{(c)})}^{(k+1)}]) \tag{4.7}$$

where $f_z : \mathbb{R}^{2\mathcal{D}} \to \mathbb{R}^{\mathcal{D}}$ is a linear transformation.

4.4.3 Inference and Learning

To identify the correct answer $a_{correct} \in \mathcal{A}$ with a QA example $(c, q, \mathcal{A})$, we compute the probability $p(a|c, q)$ for each answer choice with its multimodal semantic knowledge from scene-graph, concept-graph, QA-context node, and QA-concept node. With various node representations on the *L-th* $(L = k+1)$ layer updated by GNN modules (shown in Fig. 4.2), we obtain pooling representations $\boldsymbol{h}_{(s)}^{(k+1)}$ and $\boldsymbol{h}_{(c)}^{(k+1)}$ of scene-graph and concept-graph and then concatenate with QA-context node and

QA-concept node representations. Finally we calculate $p(a|c,q)$ by

$$\boldsymbol{h}_a^{(k+1)} = [\boldsymbol{h}_{(s)}^{(k+1)}||\boldsymbol{h}_{(c)}^{(k+1)}||\boldsymbol{h}_{(p)}^{(k+1)}||\boldsymbol{h}_{(z)}^{(k+1)}], \tag{4.8}$$

$$\text{logit}(a) = f_c(\boldsymbol{h}_a^{(k+1)}), \tag{4.9}$$

$$p(a|c,q) = \text{softmax}_a(\text{logit}(a)) \tag{4.10}$$

where $\text{logit}(a)$ indicates the confident score of answer choice a, $f_c : \mathbb{R}^{4\mathcal{D}} \rightarrow \mathbb{R}^1$ is a linear transformation that maps the concatenation of representations to a scale. We normalize it across all answer choices using the softmax function. For the training process, we apply the cross entropy loss to optimize the *VQA-GNN* model end-to-end.

4.5 Experiments

4.5.1 Experiment Setup

Visual Commonsense Reasoning (VCR). We evaluate VQA-GNN on VCR [85]. It contains 290k pairs of questions, answers, and rationales, over 110k unique movie scenes. VCR consists of two tasks: visual question answering (Q→A), and answer justification (QA→R). Each question in the dataset is provided with four candidate answers. The goal of (Q→A) is to select the best answer, while the goal of (QA→R) is to justify the given question-answer pair by picking the best rationale out of the four candidates. We joint train *VQA-GNN* on Q→A and QA→R, with a common LM encoder, the multimodal semantic graph for Q→A, a concept graph retrieved by giving question-answer pair with a rationale candidate for QA→R. We use a pretrained RoBERTa Large model to embed the QA-context node and finetune its all parameters with the multimodal GNN for 50 epoch by using learning rates 1e-5 and 1e-4 respectively. We set the number of layers ($L = 5$) of VQA-GNN and use AdamW [80] optimizer to minimize the loss. We use a linear warmup of the learning rate over the *15-th* epoch, with a cosine decay thereafter to *0*.

GQA dataset. It contains open-ended questions (1.5M questions correspond to 1,842 answer tokens), along with 110K scene graphs and the semantic functional programs to offer unambiguous instructions [93]. We only use questions without giving a semantic feature program that limits the development of the model's reasoning abilities in a more practical setting. We define the question as the context node (node q) to fully connect visual and textual scene graphs (SG) respectively to structure multimodal semantic graphs. The node q is embedded with a pretrained RoBERTa large model, and we initialize object nodes' representations in visual SG using official object features, object nodes in textual SG by concatenating GloVe [123] based word embedding of the object name and attributes. Different from the training target of VCR, the goal of GQA is to classify the given image-question pair out of 1,842 answer classes. We finetune the node q with VQA-GNN for 50 epoch by using learning rates 2e-5 and 2e-4 respectively.

4.5.2 Performance

Evaluation on VCR dataset

Comparison with state-of-the-art methods. We compared *VQA-GNN* with state-of-the-art methods on the VCR test set in Tab. 4.1. Compared with the unidirectional fusion method *SGEITL+VLBERT* that can boost multimodal transformer model VLBERT by incorporating visual scene graphs, *VQA-GNN* is a multimodal GNN-based bidirectional fusion method built on the multimodal semantic graph. Both were not pretrained on the large-scale dataset. *VQA-GNN* improves SGEITL+VLBERT on the Q→AR metric by **3.2%**, and further reduces over **11M** training parameters. We think that the structured multimodal semantic graph provides much more commonsense knowledge related to QA and original image than SGEITL, and the multimodal GNN-based bidirectional fusion method works much better on unifying unstructured and structured multimodal knowledge than multimodal transformer models. Moreover, since we retrieve commonsense knowledge from structured multimodal semantic graphs directly, *VQA-GNN* is a cost-effective approach compared to multimodal transformer models that consume much GPU resources to learn commonsense knowledge with large parameters.

We also demonstrate the effectiveness of *VQA-GNN* by comparing it with state-of-the-art multimodal transformer models that were pretrained across text and images and were finetuned on the VCR dataset. As shown in Tab. 4.1, the larger image caption data and parameters, the higher performance the model can achieve. In contrast, *VQA-GNN* trained with VCR dataset with $290K$ image-caption pairs performs similarly to UNITER-L that requires over **32x** larger image-caption data than us in the pretraining process. These results suggest that *VQA-GNN* obtaining

Model	# Image-caption in pretraining	Parameters	Structured knowledge	Test Acc.(%)		
				Q→A	QA→R	Q→AR
ViLBERT [24]	3.3M	**221M**	No	73.3	74.6	54.8
VLBERT-L [25]	3.3M	383M	No	75.8	78.4	59.7
SGEITL+VLBERT [86]	290k	$\geq$ 383M	Yes	76.0	78.0	59.6
UNITER-(B/L)[84]	9.5M	154M/378M	No	75.0/77.3	77.2/80.8	58.2/62.8
ERNIE-ViL-(B/L) [83]	3.8M	212M/533M	No	77.0/79.2	80.3/83.5	62.1/66.3
VQA-GNN (Ours)	**290k**	372M	Yes	77.9	80.0	62.8
MERLOT [27]	180M	223M	No	80.6	80.4	65.1
RESERVE-(B/L) [28]	1B	200M/644M	No	79.3/84.0	78.7/84.9	62.6/72.0
RESERVE-L + **VQA-GNN (Ours)**	1B	1B	Yes	**85.3**	**86.9**	**74.3**

Table 4.1: Accuracy scores for VCR test set. *VQA-GNN* outperforms *SGEITL+VLBERT* model on Q→AR metric by **3.2%**, and achieves competitive accuracy with SOTA methods, which have a close number of parameters but SOTA methods require a large amount of image caption data in pre-training process (over 13x larger than our model), *e.g.*, *"UNITER-L"*, *"ERNIE-ViL-B"*, *"RESERVE-B"*. Moreover, *"RESERVE-L+VQA-GNN"* outperforms *RESERVE-L* by **2.3%** on Q→AR metric.

structured context knowledge inferred from image-level and concept-level knowledge sources is as effective as the pretraining process for previous methods. Moreover, *VQA-GNN* can further enhance RESERVE-L performance on both Q→A and QA→R and finally improves the score by **2.3%** on Q→AR metric. As correcting some questions requires the model to understand commonsense knowledge related to image context and have good reasoning ability, it is difficult for multimodal transformer methods including RESERVE-L. On the other hand, *VQA-GNN* not only structures a joint semantic graph to provide commonsense knowledge related to image context but also has a good reasoning ability thanks to its multimodal GNN module. Additionally, in the supplementary material, we detail the results compared to baselines pretrained only with the VCR dataset, as well as the evaluation of different question types.

Effectiveness of the multimodal semantic graph. To further study the behavior of modules in the multimodal semantic graph, and quantitatively evaluate pretrained models used in this work (*e.g.*, RoBERTa-L, scene-graph[scene graph generator], concept-graph[conceptNet KG]), we report the performance of using different node representations in Tab. 4.2. We respectively build classification models by applying Node p and Node z to get their validation accuracy on Q→A subtask. The scene-graph structured by connecting Node p and Node z with extracted visual scene graph improves over 25% on average of these two nodes. In terms of concept-graph, it is structured by connecting Node z with retrieved conceptual triplets from

ConcepNet KG, improving Node z's performance by 15.2%. We further compare *VQA-GNN* on "scene-graph + concept-graph" w/ and w/o Node p, and the result shows that including Node p can further improve the performance by 2%. We believe that the Node p representing global visual knowledge associated with the correct answer is able to pass visual commonsense knowledge to the multimodal semantic graph, and it is effective besides employing ConcepNet KG to obtain textual commonsense knowledge [89].

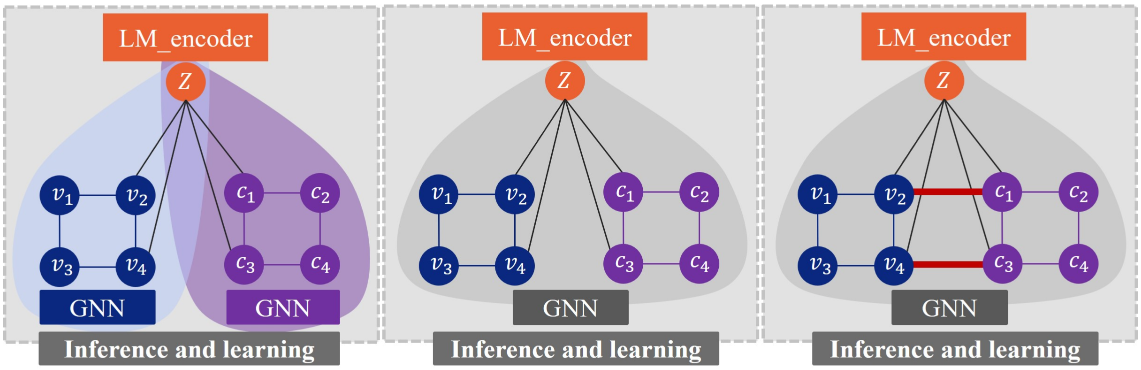
Model	Val Acc.(%) (Q→A)
Node p (Vinvl)	43.5
Node z (RoBERTa-L)	53.8
concept-graph	69.0
scene-graph	73.7
concept-graph + scene-graph (w/o node p)	75.1
concept-graph + scene-graph (w/ node p)	**77.1**

Table 4.2: All modules in the multimodal semantic graph help boost the final performance. Here, "scene-graph" includes node z and node p, "concept-graph" includes node z.

Model	Val Acc.(%) (Q→A)
Ablation 1 (single GNN)	73.0
Ablation 2 (single GNN w/ cross-modal edges)	70.6
VQA-GNN (two modality-specialized GNNs)	**75.1**

Table 4.3: Ablation 1 and Ablation 2 indicate a single GNN on the multimodal semantic graph w/o and w/ direct cross-modal edges, respectively (Fig. 4.4). VQA-GNN with two modality-specialized GNNs on the multimodal semantic graph achieves the best score.

Analysis of the multimodal GNN method. To analyze the effect of the multimodal GNN method on mitigating the multimodal gap in performing inter-modal message passing, we compared the final VQA-GNN with two single GNNs built on multimodal semantic graphs with and without direct cross-modal edges in Fig. 4.4. As the results of VCR validation set shown in Tab. 4.3, the final VQA-GNN built with the multimodal GNN on the multimodal semantic graph improves the accuracy of both ablative architecture by over **2%**. We believe that the multimodal GNN built by two modality-specific GNNs can effectively avoid directly aggregating nodes from scene-graph and concept-graph to alleviate the modality gap. As a result, the



(a) Final VQA-GNN: combination of two *modality-specialized* GNNs on the multimodal semantic graph

(b) Ablation 1: single GNN on the multimodal semantic graph

(c) Ablation 2: single GNN on the multimodal semantic graph with direct cross-modal edges

Figure 4.4: Ablation architectures. We find that our final VQA-GNN architecture with two modality-specialized GNNs overcomes the representation gaps between modalities (§4.4.1).

inter-modal message passing can be improved. We further explored the aggregation process for some node samples to demonstrate why the two ablation architectures fail to alleviate the modality gap. Here, $m^{(k)}_{\mathcal{N}(u)}$ represents the aggregated messages from all neighbors of node u at the k-th iteration.

$$m^{(k)}_{\mathcal{N}(u)} = \text{Aggregate}^{(\text{k})}(u^{(k)}, \forall v \in \mathcal{N}(u)) \tag{4.11}$$

where $\mathcal{N}(u)$ denotes a set of neighborhood nodes of the node u, and k denotes the iterations of $m^{(k)}_{\mathcal{N}(u)}$.

For (c) Ablation 2 in Fig. 4.4, we assume that node v_2 is connected with node c_1 as both represent the same notion. However, their feature vectors are distributed in different modality domains and affect the aggregation process. We show the neighborhood nodes of QA-context node z, visual node v_2 and concept node c_1 are follows:

$$\mathcal{N}(z) = \{v_2, v_4, c_1, c_3\} \tag{4.12}$$

$$\mathcal{N}(v_2) = \{z, v_1, v_4, c_1\}; \mathcal{N}(c_1) = \{z, c_2, c_3, v_2\} \tag{4.13}$$

where their neighborhood nodes include heterogeneous nodes from different modality domains.

For (b) Ablation 1 in Fig. 4.4, the neighborhood nodes of QA-context node z, visual node v_2 and concept node c_1 are follows:

$$\mathcal{N}(z) = \{v_2, v_4, c_1, c_3\} \tag{4.14}$$

$$\mathcal{N}(v_2) = \{z, v_1, v_4\}; \mathcal{N}(c_1) = \{z, c_2, c_3\} \tag{4.15}$$

Compared with (c) Ablation 2, node c_1 and node v_2 are removed from the neighborhood nodes of v_2 and c_1 which helped improve the performance of (c) Ablation 2 by 2.4%. However, it is limited by the QA-context node z that aggregates messages across scene-graph and concept-graph. Although QA-context node z is a pretrained LM that can be finetuned on multimodal domains, it is more difficult to adapt to two modalities (Eq. 4.14) than to a single modality (Eq. 4.16). In contrast, the multimodal GNN method is designed by introducing two GNNs for each modality. We perform aggregation for QA-context node z for each modality so that the pretrained LM is finetuned on a single modality to alleviate the modality gap. The neighborhood nodes of QA-context node z, visual node v_2 and concept node c_1 are follows:

$$\mathcal{N}(z)^{(m1)} = \{v_2, v_4\}; \mathcal{N}(z)^{(m2)} = \{c_1, c_3\} \tag{4.16}$$

$$\mathcal{N}(v_2) = \{z^{(m1)}, v_1, v_4\}; \mathcal{N}(c_1) = \{z^{(m2)}, c_2, c_3\} \tag{4.17}$$

where $m1$ and $m2$ indicate two message passing methods for each modality.

Evaluation on GQA dataset

Comparison with baselines. We also compared *VQA-GNN* with baseline models on GQA dataset, under the realistic setup of not using the annotated semantic functional programs (see §4.5.1). As the results shown in Tab. 4.4, our model achieves validation accuracy of 58.9% for visual SG and 87.9% for textual SG. Compared with SGEITL [86] and GCN [99] which are unidirectional fusion methods, our method performs bidirectional fusion to unify unstructured and structured knowledge, and improved the reasoning ability of SGEITL by **5.6%** and GCN by **2.2%**. Moreover, by inter-connecting the visual and textual SG, our method achieves validation accuracy of **90.3%** and further suggests its efficacy in performing inter-modal message passing.

Ablation study on the bidirectional fusion. To fairly study the effect of bidirectional fusion for improving concept-level reasoning, we evaluated the performance

Model	Visual SG	Textual SG	Val Acc.(%)
SGEITL[86]	✓		53.3
CFR[124]	✓	✓	73.6
GCN[99]		✓	85.7
VQA-GNN	✓		58.9
		✓	87.9
	✓	✓	**90.3**

Table 4.4: Accuracy scores on the GQA validation set. All models are trained under the realistic setup of not using the annotated semantic functional programs.

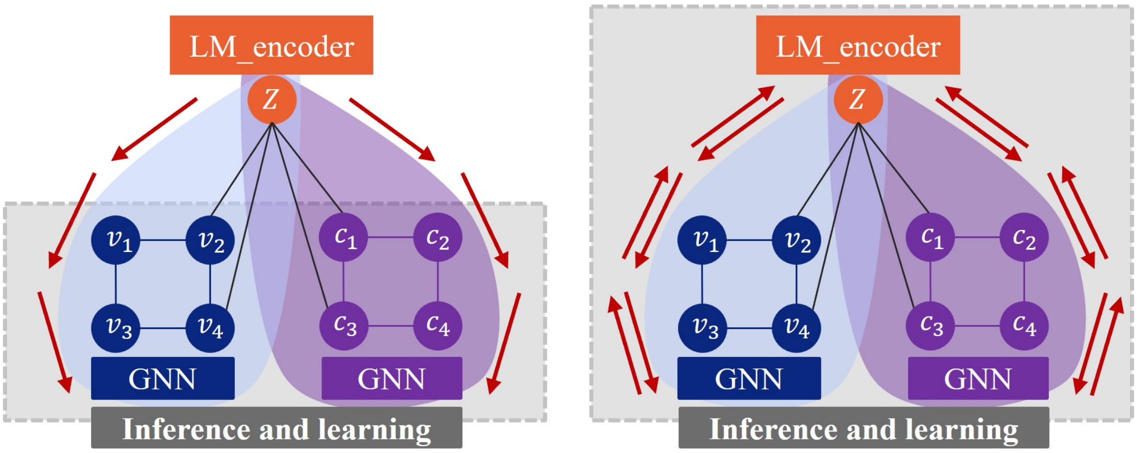
Method	Val Acc.(%) ↑	Inference time (ms) ↓
Average pooling	62.3 (±0.40)	**5.2**
Unidirectional fusion	86.3 (±0.01)	8.6
Bidirectional fusion **(ours)**	**90.3** (±0.03)	5.5

Table 4.5: Ablation results on the effect of our proposed bidirectional fusion for GQA.

of *VQA-GNN* with and without structured multimodal knowledge-enhanced question representations. We show their difference in Fig. 4.5, compared with the unidirectional fusion, the bidirectional fusion approach is able to utilize the message aggregated from scene-graph and concept-graph in node z to predict the correct answer. It facilitates the joint reasoning ability of VQA-GNN in capturing bidirectional interactions between unstructured node z and structured multimodal semantic graph. As a result in Tab. 4.5, the bidirectional fusion approach further improved the performance of the unidirectional fusion approach by **4%**. We also compared our approach with an average pooling method that simply averages all node representations. We indeed find that this ablation performs significantly worse than others, which suggests that our approach can capture special relationship information between different nodes but average pooling cannot.

4.5.3 Comparison with baselines pretrained only on VCR dataset

We compared VQA-GNN and multimodal transformer models in Tab. 4.6 which were only trained on VCR dataset (290K instances), as reported in SGEITL paper [86]. SGEITL is an add-on module that can boost multimodal transformer models



(a) Unidirectional fusion: prediction with scene-graph and concept-graph

(b) Bidirectional fusion: prediction with the multimodal semantic graph

Figure 4.5: Illustration of two knowledge fusion methods: our proposed bidirectional fusion v.s. the unidirectional fusion baseline.

(UNITER, VLBERT) by incorporating finetuned visual scene graph with multimodal transformer models. Compared with SGEITL, VQA-GNN is a GNN-based method built on the structured multimodal semantic graph. As shown in Tab. 4.6, VQA-GNN improves over SGEITL+VLBERT on the Q→AR metric by **4%** for the validation set, and further suggests the efficacy of VQA-GNN on the well-structured multimodal semantic graph.

Model	Parameters	Val Acc.(%)		
		Q→A	QA→R	Q→AR
VLBERT-L	383M	72.9	75.3	54.9
UNITER-L	378M	73.4	76.0	55.8
ERNIE-ViL-L	533M	74.1	76.9	56.9
SGEITL+UNITER	$> 378M$	74.8	76.8	57.4
SGEITL+VLBERT	$> 383M$	74.9	77.2	57.8
VQA-GNN(ours)	**372M**	**77.1**	**80.0**	**62.1**

Table 4.6: All models are trained only on the VCR dataset. Compared to the “SGEITL+VLBERT” model that inputs a visual scene graph to VLBERT network, VQA-GNN applied to a well-structured multimodal semantic graph improves accuracy on Q→AR metric by over **4%**.

4.5.4 Comparison results on different question types

We studied the performance of VQA-GNN in different question types and compared it with a strong baseline model RESERVE-L in Tab. 4.7. VQA-GNN outperforms RESERVE-L in some question types such as "Will", "Have", and "Can/Should", and we consider that some questions require the model to understand commonsense knowledge related to image context and have good reasoning ability. Hence, the model "RESERVE-L+VQA-GNN" boosted the performance of RESERVE-L.

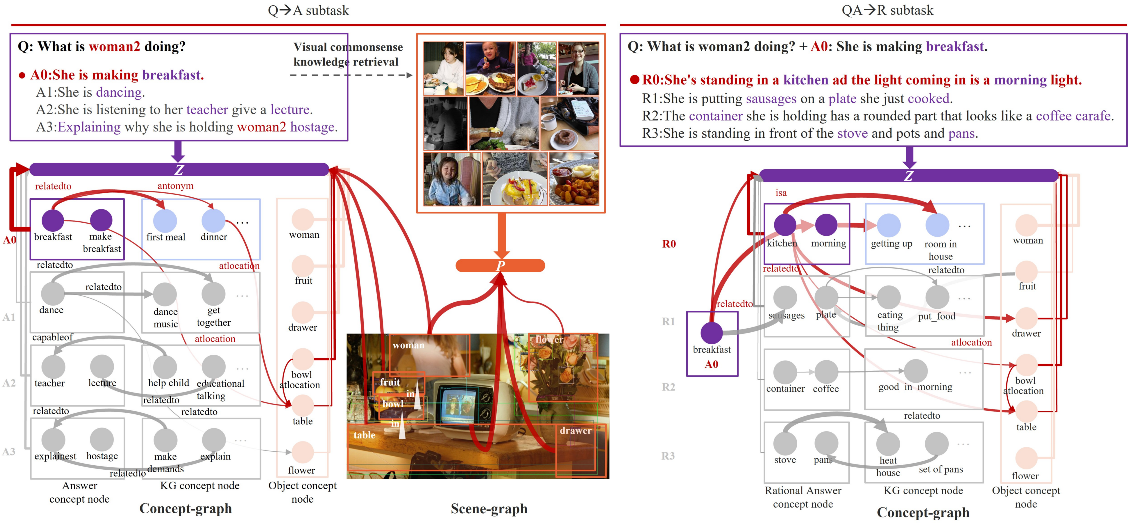
Question type	Val Acc.(%) (Q→A)		Val Acc.(%) (QA→R)	
	VQA-GNN	RESERVE-L	VQA-GNN	RESERVE-L
Why	73.2	**78.6**	81.8	**84.8**
What	79.1	**85.7**	80.0	**85.2**
Where	77.9	**87.7**	76.7	**86.0**
Who	89.4	**91.3**	77.1	**85.0**
When	77.8	**94.4**	**100**	100
Which	**88.9**	88.9	81.5	**87.0**
Do	**81.6**	81.6	73.5	**82.5**
Will	**86.2**	83.8	**82.7**	82.3
Have	**92.9**	91.4	**87.1**	82.9
If	89.2	**92.3**	**96.9**	95.4
Can/Should	**93.3**	88.5	**87.5**	84.6

Table 4.7: Comparison results on the different question types. VQA-GNN performs better than RESERVE-L for "Will", "Have" and "Can/Should" question types.

4.6 Interpretability

To interpret how *VQA-GNN* reason a correct answer based on a structured multimodal semantic graph, we show the reasoning process on Q→A and QA→R subtasks of VCR respectively in Fig. 4.6 by using a validation sample that is given a correct answer on both Q→A and QA→R subtasks by *VQA-GNN*.

***Q→A* subtask.** We trace high attention weights from two directions: **d1**: QA-context node **Z** → **A**nswer nodes (purple) → **KG** concept nodes (blue) → **O**ject concept nodes (pink); **d2**: QA-concept node **P** → **SG** object nodes (orange) → **Z**. At the **d1**, **Z** pays more attention to **A** nodes "breakfast" and "make breakfast" in answer "A0" choice than nodes in other choices, "breakfast" attends to both **KG**


Q→A subtask

Q: What is woman2 doing?

- A0:She is making breakfast.

A1:She is dancing.
A2:She is listening to her teacher give a lecture.
A3:Explaining why she is holding woman2 hostage.

QA→R subtask

Q: What is woman2 doing? + A0: She is making breakfast.

- R0:She's standing in a kitchen ad the light coming in is a morning light.

R1:She is putting sausages on a plate she just cooked.
R2:The container she is holding has a rounded part that looks like a coffee carafe.
R3:She is standing in front of the stove and pots and pans.

Figure 4.6: Interpreting *VQA-GNN*'s reasoning process across multimodal knowledge domains by indicating attention weight of the relationship between nodes. Arrows indicate the direction of the relationship, and darker and thicker edges indicate higher attention weights. The red color highlights the message passing routine for reasoning the correct answer and the gray color indicates the opposite.

node "first meal" and **O** node "table", **O** node "table" further attends to **O** node "bowl", and both strongly attend to **Z**. **A** node "breakfast" bridges the reasoning between **Z** and **O** "table" at the concept-level. Besides with **d1**, we also track the attention weights from **d2**, **Z** strongly attends to **SG** nodes "table", "drawer" and "woman", all nodes attend to **Z**, which reveals image-level semantic knowledge of **SG** nodes "table", "drawer" and "woman" are all essential for reasoning "**A0:** she is making breakfast" correct. These two reasoning paths demonstrate that *VQA-GNN* is an inoperable method that can give a reasonable explanation to each choice with our well-structured multimodal semantic graph, also suggest that our multimodal semantic graph is capable of unifying unstructured (*e.g.*, QA-context node and QA-concept node) and structured (*e.g.*, scene-graph and concept-graph) multimodal knowledge.

***QA→R* subtask.** We trace reasoning path for the rational answer **R0** on concept-graph. There are two reasonable directions: **Z** → "breakfast" → "morning" → "getting up", and **Z** → "kitchen" → "drawer", "bowl", "table". Both of them show strong attention between QA text and **R0**, compared to the attention direction for **R1** indicating that "breakfast" also strongly attends to "sausages" and "plate" attends to "fruit", however, "fruit" weakly attends to **Z**. As a result, *VQA-GNN* can select a rational answer, and suggest its interpretability on $QA \rightarrow R$ subtask. In addition, we noted that our method has close reasoning paths that attend to the image context of “bowl”, “table” and “drawer” on both $Q \rightarrow A$ and $QA \rightarrow R$ subtasks. Hence, we consider that our method has strong reasoning ability across multimodal knowledge domains.

4.7 Conclusion

We proposed a novel visual question answering method, *VQA-GNN*, which unifies unstructured and structured multimodal knowledge to perform joint reasoning of the scene. In the evaluation of two challenging VQA tasks (VCR and GQA), our method substantially outperforms existing models without pretraining using massive image-caption data under the same training setting, our method outperforms strong baseline VQA methods by **3.2%** on VCR (Q-AR) and **4.6%** on GQA, suggesting its strength in performing concept-level reasoning. Ablation studies further demonstrate the efficacy of the bidirectional fusion and multimodal GNN method in unifying unstructured and structured multimodal knowledge.

Part 2: Achieving effective multimodal transfer learning

Chapter 5

LDNN: Linguistic Knowledge Injectable Deep Neural Network for Group Cohesiveness Understanding

In Part 1, we explored the effects of varying representation distributions across different modalities and developed methods to enhance the effectiveness of multimodal fusion. In the last chapter, we proposed a new bidirectional fusion to enable the expressive integration of multimodal structured graph knowledge and unstructured knowledge in the pretrained language models. Specifically, our proposed method VQA-GNN achieved fine-grained video understanding for interpretable multimodal reasoning by unifying commonsense knowledge within visual scenes.

In Part 2, we focus on a new problem setting for developing an efficient modality-agnostic multimodal system—an unimodal system that can achieve competitive performance with a multimodal system when only provided with an unimodal signal input. In this chapter, we propose a simple approach called linguistic knowledge injectable deep neural network (LDNN) to build a visual model (visual LDNN) that can automatically associate the linguistic knowledge hidden behind images when provided with a single visual modality as input.

5.1 Introduction

Group cohesion is one of the essential characteristics of group activities and reflects the level of intimacy people feel with each other [125, 126, 127]. The greater the level of group cohesiveness, the closer the relationships between the people. Thus, the prediction of group cohesiveness is an important task for achieving a dialog robot that can facilitate human communication [128].

Due to many features correlated to group cohesiveness that exist in the image, it is an intuitive way to extract those visual features to predict the levels of cohesion. There have been some studies that have attempted to extract visual features such as scene, skeleton, especially the facial expression features of group members using pretrained DNN-based models, and apply these in an attempt to group cohesiveness prediction [129, 130, 131]. Although these visual features are related to group cohesiveness, they cannot represent the group cohesiveness concept semantically like the language representations. As shown in Fig. 5.1, the language representations for "A young mother is comforting her daughter" contain much more knowledge of group cohesiveness than visual representations extracted from image pixels such as "a little girl", "one woman". Furthermore, visual feature extraction consumes relatively higher computational time, which makes it hard to meet the demands of practical applications.

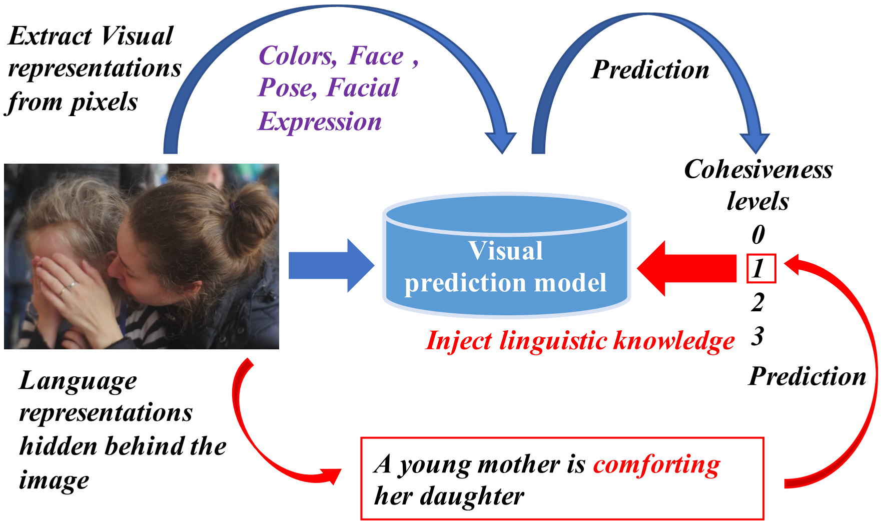

Extract Visual representations from pixels
Colors, Face , Pose, Facial Expression
Prediction
Cohesiveness levels
0
1
2
3
Visual prediction model
Inject linguistic knowledge
Prediction
Language representations hidden behind the image
A young mother is comforting her daughter

Figure 5.1: A conceptual diagram illustrating the injection of linguistic knowledge into a visual prediction model.

Exploring ways to obtain the linguistic knowledge of group cohesiveness hidden behind images makes sense to improve performance. Image captioning is a multimodal translation task that maps visual domain-specific features into language domain-specific features by implementing an encoder-decoder framework [132]. It can be used to get visual features that reflect the language features, however, these language features cannot represent group cohesiveness directly. On the other hand, multimodal representation learning aims to generate informative representations by integrating different modal features and has been widely applied to emotion recognition tasks [133] and visual dialog tasks [134], etc. Although it is a useful way to gain both visual and language representations correlating group cohesiveness, language information is essential for practical use.

Inspired by the fact that humans intuitively associate linguistic knowledge accumulated in the brain with the visual images [135], we propose a linguistic knowledge injectable deep neural network (LDNN) for constructing a visual model (visual LDNN) that can associate related linguistic knowledge of group cohesiveness hidden behind images without any language information at inference time. LDNN is composed of two components, a visual encoder, and a language encoder, to apply multimodal domain adaptation and linguistic knowledge transition mechanisms. We train the LDNN by adding descriptions to the training and validation set of the Group AFfect Dataset 3.0 (GAF 3.0) [136] and test the visual LDNN without any description. Comparing visual LDNN with various fine-tuned DNN models and three state-of-the-art models in the test set, the results demonstrate that the visual LDNN not only improves the performance of the fine-tuned DNN model leading to an MSE very similar to the state-of-the-art model but is also a practical and efficient method that requires relatively little preprocessing. Further ablation studies are convinced that LDNN is an effective method to inject linguistic knowledge into visual models. The contributions of this paper can be summarized as follows:

(i) We propose LDNN which can transform the linguistic knowledge distilled from a language model and transfer it into a single visual model.

(ii) We train a linguistic knowledge injected visual model using language and visual modal information only in the training phase and single visual modal information in the inference phase.

(iii) We expand an existing dataset of GAF 3.0 by adding a description to each video data and show performance comparable to state-of-the-arts using multimodal information as single modal information.

5.2 Related works

Group cohesiveness understanding from video data is becoming a standard way by employing advanced artificial intelligence techniques such as deep neural network (DNN)-based methods [125, 126, 127]. Many studies have proven that DNN-based methods can perform extremely well in the field of image understanding, not only in relation to coarse-grained tasks such as the MNIST task but also in fine-grained tasks such as facial expression recognition tasks [137]. These works [129, 130, 131] propose DNN-based methods for the prediction of group cohesiveness in images by adopting a common strategy, namely, extracting as many visual features as possible including scenes, skeletons, and faces utilizing pretrained DNN models. Although these visual features correlate with group cohesiveness, it cannot be guaranteed that all features can always be extracted from any image, such as when someone's face is hidden by others in some images. In addition, the linguistic knowledge that describes the images helps to represent the interpersonal intimacy of group members, but this matter is not considered in these studies.

Language scaffolds concept knowledge in humans, helping them to acquire abstract concepts [138]. Analyzing image descriptions will help to obtain complex representations that facilitate the understanding of interpersonal intimacy. With recent advances in natural language processing (NLP), human language can be easily transformed into a high-dimensional vector with embedded linguistic knowledge [21]. Bidirectional encoder representations from transformers (BERT) is a state-of-the-art language model for NLP, by which a pretrained model can be fine-tuned to produce state-of-the-art results in a wide range of NLP tasks [72], such as question answering (SQuAD) [139], machine reading comprehension [140], and visual commonsense reasoning [25]. We fine-tune a pretrained model of BERT to obtain a high-dimensional vector from the image description to represent linguistic knowledge of group cohesiveness.

It makes sense to merge much more information to help understand complex concepts. Many tasks focus on how to integrate multimodal information that results in high recognition performance. The work [59] proposes a multimodal fusion method for the human emotion recognition task. It takes features extracted by the visual and audio encoders as the inputs and applies attention mechanisms that highlight important modal features to achieve higher scores than single modalities.

[133, 141] proposed an emotion recognition method that attempts to achieve human-like emotion recognition by integrating features extracted from visual, audio, and linguistic data. The visual question answering (VQA) [12] task is a much more challenging multimodal task that requires not only extracting the important features from the image and question text but also exploring the interaction between visual and linguistic features. The work [142] proposes a method of applying a co-attention mechanism to dynamically learn the interaction between objects in images and question text. However, these methods need to provide multimodal information to generate meaningful representations and cannot be applied in practice without providing the linguistic information hidden behind images. Image captioning is a task that generates image descriptions by learning the mapping relationships between image pixels and descriptions [143, 144, 145]. Although existing methods can produce mapping features based on the objects present in the image, it is not enough to represent linguistic knowledge that is hidden behind images. Thus, an image captioning method is not the best choice to apply to extend the learning of high-level representations to predict cohesive levels.

5.3 Method

We propose an LDNN based on the concept that humans unconsciously associate linguistic knowledge accumulated in the brain with visual images [135]. We aim to construct a visual model for group cohesiveness understanding that can consider related linguistic knowledge hidden behind images. As shown in the top part of Fig. 5.2, an LDNN consists of a visual encoder and a language encoder for processing raw images and image descriptions respectively, and applies two mechanisms: domain adaptation and linguistic knowledge transition. As shown in the bottom of Fig. 5.2, the multimodal domain adaptation mechanism is used to map visual and language representations into a common vector space to influence the process of linguistic knowledge transition; the linguistic knowledge transition mechanism is designed to inject linguistic knowledge into the visual model.

LDNN is an end-to-end architecture in which the visual model and language model are jointly trained corresponding to the cohesive levels. As shown in the top part of Fig. 5.2, the LDNN takes two inputs (an image and its description) for training, the visual model (visual LDNN) is built with an adaptation layer and a knowledge layer following a visual encoder, and the language model is built with an

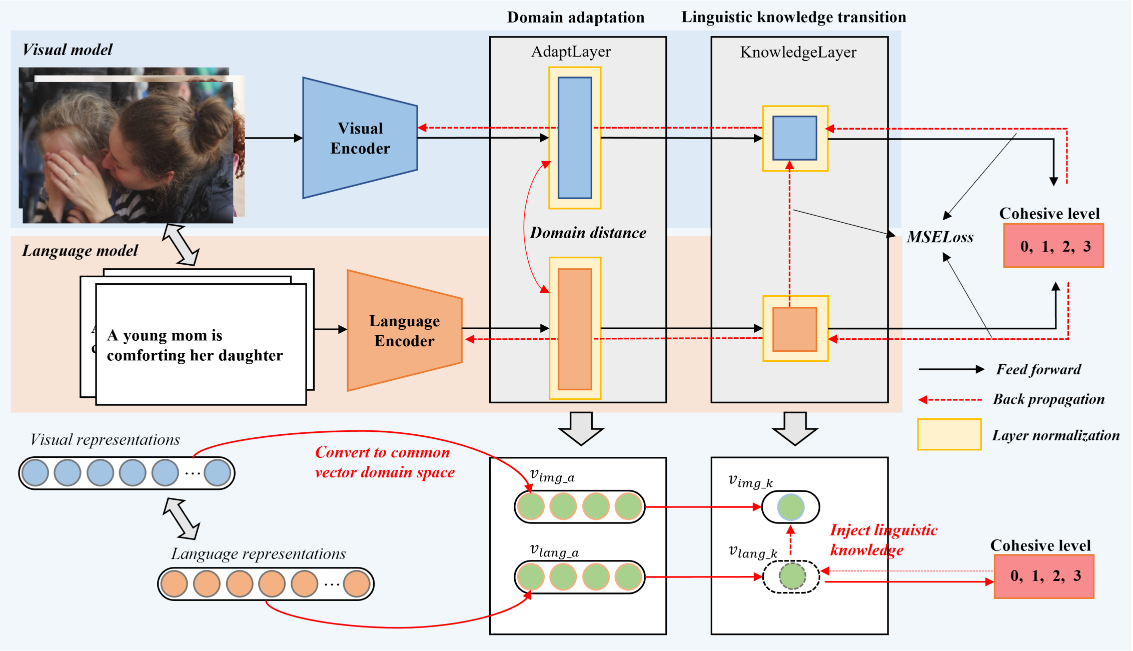


Figure 5.2: Architecture for training LDNN. The visual model and the language model are jointly trained to inject linguistic knowledge into the visual model. The visual encoder is a pretrained DNN model for extracting visual representations. The Language encoder is a BERT-based pretrained model for extracting language representations. The bottom part of this figure shows the learning process of injecting linguistic knowledge.

adaptation layer and a knowledge layer following an language encoder. To train the visual LDNN that can obtain both visual and linguistic knowledge of the cohesive levels simultaneously, we provide the learned linguistic knowledge as another target of the visual model. Thereby, linguistic knowledge can be injected into the visual LDNN.

As shown in Fig. 5.3, we only utilize a visual LDNN to predict the cohesive levels of an image. Even though we do not provide language information hidden behind the image, a visual LDNN can associate linguistic knowledge that has been extracted from the language model through training.

5.3.1 Visual Encoder and Language Encoder

Visual encoder: The visual encoder extracts the feature map output by DNN models pretrained on the ImageNet dataset. The input image I is converted to a 2048-dimensional vector v_{img}. We apply seven DNN models with performance variations as the visual encoders of the visual LDNN, which are AlexNet, VGG11,

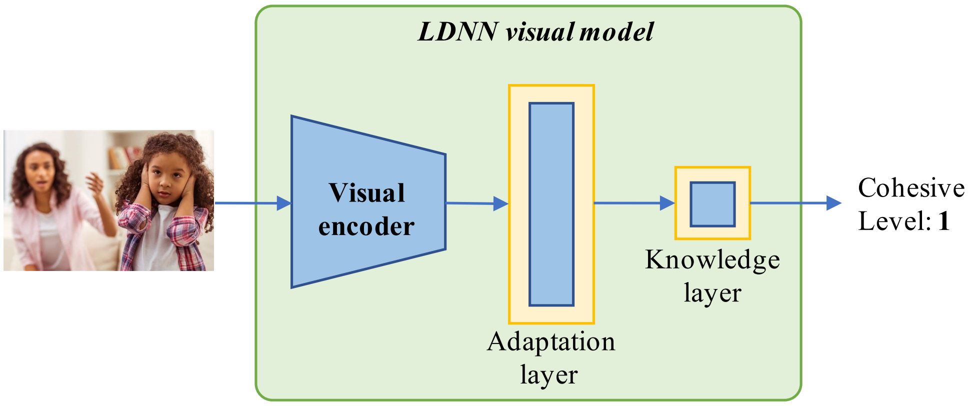


Figure 5.3: LDNN visual model used for predicting the cohesive level of an image.

VGG16, ResNet18, ResNet50, DenseNet161 and SENet154 [42, 16, 17, 146]. In section 4, we evaluate the performance of the visual LDNN and discuss how linguistic knowledge affects the performance of the visual LDNN by applying various visual encoders.

$$v_{img} = \text{VisualEncoder}(I) \tag{5.1}$$

Language encoder: The language encoder is composed of a BERT embedding layer ("BertLayer") and a self-attention layer ("SelfLayer"), which have the function of transforming the image description into an embedding vector v_{lang} [21]. In the "BertLayer", we use a BERT model pretrained on the Japanese Wikipedia corpus to extract word embedding vector by giving the image description [72]. We input the words $\{w_1, ..., w_n\}$ of the image description to the "BertLayer", and output an embedding metric V_{lang} with the size of (n, d). Here, n and d denote the number of words and the dimension size of the word vector respectively. In the "SelfLayer", to emphasize important words in the image description, we calculate attention energy $Attn\_ene$ with a size of $(n, 1)$ by inputting V_{lang} to a feed-forward sub-layer composed of a single neuron with a softmax function. Here, we apply the softmax function to ensure all the attention energy of words sum up to 1. The output of the "SelfLayer" is an d-dimensional embedding vector v_{lang} that is the sum of the element-wise product of $Attn\_ene$ and V_{lang}.

$$V_{lang} = \text{BertLayer}(\{w_1, ..., w_n\}) \tag{5.2}$$

$$v_{lang} = \text{SelfLayer}(V_{lang}) \tag{5.3}$$

5.3.2 Domain adaptation and linguistic knowledge transition

Domain adaptation: While distinct modal features can represent the same target, the domain disparities between these features pose challenges in their direct joint processing. It makes sense to map multimodal features to a common domain space, not only for facilitating the integration of multimodal features but also for learning the correlations between multimodal features. To learn the correlations between visual and language representations and facilitate the following linguistic knowledge transition process, we build the domain adaptation mechanism to learn the mapping of visual and linguistic representations by minimizing their vector distance at the backpropagation process.

We apply two feed-forward sub-layers ("AdaptLayer") to LDNN. These sublayers take the output of visual and language encoders as input and process the visual adaptation vector v_{img_a} and language adaptation vector v_{lang_a}. We calculate the pairwise distance between two adaptation vectors and define it as the loss function $\mathcal{L}_a$. Here, the "LayerNorm" is a normalization sublayer followed by the "AdaptLayer". We minimize the $\mathcal{L}_a$ to process domain adaptation at the backpropagation process.

$$v_{img\_a} = \mathrm{LayerNorm}(\mathrm{AdaptLayer}(v_{img})) \tag{5.4}$$

$$v_{lang\_a} = \mathrm{LayerNorm}(\mathrm{AdaptLayer}(v_{lang})) \tag{5.5}$$

$$\mathcal{L}_a(v_{img\_a}, v_{lang\_a}) = \left(\sum_{i=1}^{d} |v^i_{img\_a} - v^i_{lang\_a}|^2\right)^{1/2} \tag{5.6}$$

Linguistic knowledge transition: A typical fusion model requires all single modalities to train meaningful representations, in preparation for practical use. LDNN is trained on images and their descriptions to achieve a visual LDNN into which linguistic knowledge can be injected. Thus, image descriptions are not required to prepare for practical use. Inspired by the distillation methods [147] proposed to transfer pretrained knowledge to a lightweight model, the linguistic knowledge transition mechanism is designed to simultaneously obtain and inject linguistic knowledge into a visual model. Following the "AdaptLayer", we build the "knowledgeLayer", a feed-forward layer with a single unit to not only learn visual and linguistic knowledge of cohesive levels but also to transfer linguistic knowledge to the visual LDNN. We perform the following steps below to achieve linguistic knowledge transition:

1. Take the "knowledgeLayer" as two regression layers and learn the visual and linguistic knowledge of cohesive levels. We jointly minimize both regression loss $\mathcal{L}_k^{img}$ and $\mathcal{L}_k^{lang}$ by using the mean squared error (MSE) loss function defined as follow:

$$\text{MSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n}\left(\sum_{i=1}^{n} |y_i - \widehat{y}_i|^2\right) \tag{5.7}$$

   where $\mathbf{y}$ and $\hat{\mathbf{y}}$ represent a batch size n of the prediction value and ground truth label and the actual value.

$$v_{img\_k} = \text{LayerNorm}(\text{knowledgeLayer}(v_{img\_a})) \tag{5.8}$$

$$v_{lang\_k} = \text{LayerNorm}(\text{knowledgeLayer}(v_{lang\_a})) \tag{5.9}$$

$$\mathcal{L}_k^{img} = \text{MSE}(v_{img\_k}, T) \tag{5.10}$$

$$\mathcal{L}_k^{lang} = \text{MSE}(v_{lang\_k}, T) \tag{5.11}$$

   $\mathcal{L}_k^{img}$ and $\mathcal{L}_k^{lang}$ are computed by applying the outputs of the "LayerNorm" layers, v_{img_k} and v_{lang_k} to Eq. 5.7. Here, T denotes the actual cohesive level in a range of $[0, 3]$.

2. Regard the linguistic knowledge v_{lang_k} as the ground truth of the visual LDNN, transferring linguistic knowledge to the visual LDNN by forward conducting MSE loss between v_{lang_k} and v_{img_k}, which is defined as Knowledge transition loss $\mathcal{L}_k^{lang \to img}$.

$$\mathcal{L}_k^{lang \to img} = \text{MSE}(v_{lang\_k}, v_{img\_k}) \tag{5.12}$$

3. Adapt two hyperparameters of α and β to the $\mathcal{L}_a$ and $\mathcal{L}_k^{lang \to img}$ to control the effects of domain adaptation and linguistic knowledge transition mechanisms on different visual encoders. Empirically, setting α to 1 and β to 2 is reasonable for most visual encoders. Finally, we train the LDNN by minimizing $\mathcal{L}$ which is the integration of all losses.

$$\mathcal{L} = \mathcal{L}_k^{img} + \mathcal{L}_k^{lang} + \alpha * \mathcal{L}_a + \beta * \mathcal{L}_k^{lang \to img} \tag{5.13}$$

5.4 Experiment

In this section, we first introduce the dataset, experiment settings, and metrics, and then empirically compare visual LDNNs to fine-tuned DNNs and three state-of-the-art methods [129, 130, 131], aiming to demonstrate the effectiveness of LDNN. Besides, we performed two ablation experiments to ascertain why LDNN is effective in injecting knowledge of the language model into the visual model. We also assessed how effective LDNN is for various visual encoders.

5.4.1 Dataset, settings and metric

Dataset: GAF 3.0 [136] is created by web crawling various keywords related to social events such as football World Cup winners, birthday parties, and violence. It contains a total of 16,443 images, 9,300 images for training, 4,244 images for validation, and 2,899 images for testing, respectively (details in Tab. 5.1). We utilize the labels for group cohesiveness as provided by [125] except testing labels. Since the testing labels were not released, we re-annotated them with the same annotation policy as the GAF 3.0 dataset. There are samples of the GAF 3.0 shown in Tab. 5.2. The group's perceived cohesiveness is in the range from 0 to 3, where 0 represents very weak cohesion, 1 represents weak cohesion, 2 represents strong cohesion and 3 represents very strong cohesion, respectively.

Dataset	Total	0: very weak	1: weak	2: strong	3: very strong
Train	9,300	1,141	1,561	4,601	1,997
Valid	4,244	351	1,226	1,394	1,273
Test	2,899	179	606	1,717	397

Table 5.1: The size of training (Train), validation (Valid), and testing (Test) set for individual classes.

In addition, using professional annotators, we add image descriptions to the training and validation datasets for training LDNN[1]. The image descriptions are created by three annotators (one male and two females). To guarantee the quality of the annotation result, none of the annotators was allowed to refer to group cohesiveness scores and each was required to describe the image from the viewpoint of human interaction objectively. For example, the sample image at the bottom of

[1] https://github.com/wangyanan326/Additional-EmotiW-dataset

Tab. 5.2 can be annotated as “Three siblings face each other and snuggle up to each other”.

Image	Description	Cohesive level
	複数人の警察官が一人の男性を囲んでいる (Multiple police officers surround a man)	0: very weak
	少女たちが車いすに買い物袋を載せて運んでいる (Girls are carrying shopping bags on wheelchairs)	1: weak
	集団の中心で男性が少女を抱いて立っている (Guy stands in the center of the group and hugs the girl)	2: strong
	三姉弟が顔を寄せ合い寄り添っている (Three siblings face each other and snuggle up to each other)	3: very strong

Table 5.2: Samples from GAF 3.0 validation set. The description is annotated in the Japanese language and added to the training and validation dataset for training LDNN. A high score indicates strong cohesion among people shown in the image.

Settings: We select various DNN-based models pretrained on the ImageNet dataset as visual encoders. They are AlexNet, VGG11, VGG16, ResNet18, ResNet50, ResNet152, and DenseNet161, and all these models take images of size 224 × 224 as input. The output of the visual encoder is a 2048-dimensional vector extracted from the last pooling layer of the pretrained model. Regarding the language encoder, we apply Japanese and English language-based BERT models to extract word-embedded features from image descriptions. Here, the English description is translated using Google translation API from Japanese content. A single description is converted to a words embedding vector of shape (N, D), N indicates the number of words in a description and D indicates the dimension of the embedded vector. The dimension size of Japanese-based embeddings is 768 and the English-based embeddings are 1024. For the training settings, we train LDNN models by applying the Adam optimizer for 100 epochs with a batch size of 64. We also adjust the learning rate so that each model has their best score.

Metrics: We apply the Mean Squared Error (MSE) as the evaluation metric of the group cohesiveness prediction.

5.4.2 Results

We built and compared Japanese and English-based language encoders to select the better one to apply to the LDNN training. The MSE results of the validation set in Tab. 5.3 show that the Japanese-based language encoder has a much lower MSE and is superior to the English-based language encoder. The gap between the results is believed to be generated by the automatic translation of the English descriptions.

Language encoder	Dimension	MSE	
		Val	Test
English-based	1,024	0.8417	-
Japanese-based	768	**0.6899**	-

Table 5.3: Comparison results of Japanese and English-based language encoder.

We trained several LDNNs constructed using a Japanese-based language encoder and various visual encoders. We then present a comparison of the visual LDNNs with the fine-tuned DNNs in Tab. 5.4. AlexNet and ResNet18/50/152-based visual LDNNs have lower MSE than DNNs for both the validation set and the test set. VGG11/16 and DenseNet161-based visual LDNNs have comparable MSE with DNNs for the validation set and lower MSE for the test set than fine-tuned DNNs. In particular, the ResNet50-based visual LDNN has the lowest MSE in the test set and is far superior to the fine-tuned DNN. From the test results, it is clear that LDNN is an effective method for training visual models that can improve the performance of fine-tuned DNNs.

Visual Encoder	Parameters	Hyper parameter		Val(MSE)		Test(MSE)	
	(million)	α	β	fine-tuned DNN	visual LDNN	fine-tuned DNN	visual LDNN
AlexNet	58	2	2	0.8077	0.7432	0.5998	0.5167
VGG11	155	1	2	0.7336	0.7556	0.5378	0.5099
VGG16	122	1	2	0.7052	0.7067	**0.4944**	0.4862
ResNet18	11	1	2	0.6896	**0.6774**	0.4976	0.4893
ResNet50	24	1	2	0.7428	0.6892	0.5611	**0.4677**
ResNet152	58	1	2	0.7367	0.7032	0.5343	0.4678
DenseNet161	26	1	2	**0.6889**	0.6896	0.5170	0.4820

Table 5.4: Comparison results of fine-tuned DNNs and visual LDNNs on validation(Val) and test sets.

Model	Features		MSE	
	Training	Inference	Val	Test
Multi-stream Hybrid Network [129]	face, scene, skeleton, UV coordinates	same as training	0.525	0.487
Multi-task learning with regularization methods [130]	face, scene, body	same as training	**0.497**	0.525
Attention Based GRU Architecture [131]	face, scene, skeleton	same as training	0.691	**0.466**
ResNet50-based visual LDNN (ours)	scene, image description	scene	0.689	0.468

Table 5.5: Results of comparing ResNet50-based visual LDNN and state-of-the-art methods.

We also compared ResNet50-based visual LDNN with three state-of-the-art methods [129, 130, 131] under similar experimental conditions. All of the state-of-the-art results are solely trained on the training set, and not boosted by the ensemble method. As the test results in Tab. 5.5 show, the ResNet50-based visual LDNN has an MSE of 0.468, which is almost the same as the lowest MSE of 0.466 [131]. The ResNet50-based visual LDNN is trained using scene features and image descriptions instead of extracting lots of visual features such as face, scene, and skeleton, however, it only takes scene features as input during inference time. Therefore, LDNN is also a practical and efficient method that requires relatively little preprocessing. Here, even though we showed both the validation results and the test results in Tab. 5.4, 5.5, the validation set was used for tuning the parameters of our model. On the contrary, the test set is independent of the training process so the test results are much more reliable for performance evaluation [148]. As the inconsistent validation results (lower validation MSE but higher test MSE compared to other methods) and the resulting gap between the validation and test set (higher validation MSE but lower test MSE) show, it is easy to confirm that the validation results are insufficient for performance evaluation. Furthermore, we consider those results may be related to the size gap between the validation set (4,244) and the test set (2,899), some overfitting to the validation set, and the difference in the class distribution of individual datasets (details in Tab. 5.1).

5.4.3 Analysis

To prove LDNN is an effective method that can improve visual model performance by injecting linguistic knowledge of the language model into the visual model, we performed two ablation experiments to analyze the impact of the KnowledgeLayer and the Language model. In addition, we also analyzed how effective LDNN is for various visual encoders.

Ablative Study 1: We built a knowledgeLayer ablation network (KnowledgeAblationNet) by clipping the knowledgeLayer of LDNN as shown in Fig. 5.4. It was similar to image captioning methods aimed at learning the mapping between visual and language representations. We compared this network to LDNN to verify the effect of the LDNN knowledgeLayer. As the results in Tab. 5.6 show, almost all the results of visual LDNN with AdaptLayer and Knowledge Layer have much lower MSE than the visual LDNN trained with the KnowledgeAblationNet. Even though KnowledgeAblationNet could incorporate the general language representations included in the description into the visual model, these general language representations did not contain any linguistic knowledge of cohesive levels. On the other hand, the LDNN KnowledgeLayer was applied to distill the linguistic knowledge of cohesive levels that can be transferred into the visual model, and thus the trained visual LDNN could achieve state-of-the-art performance.

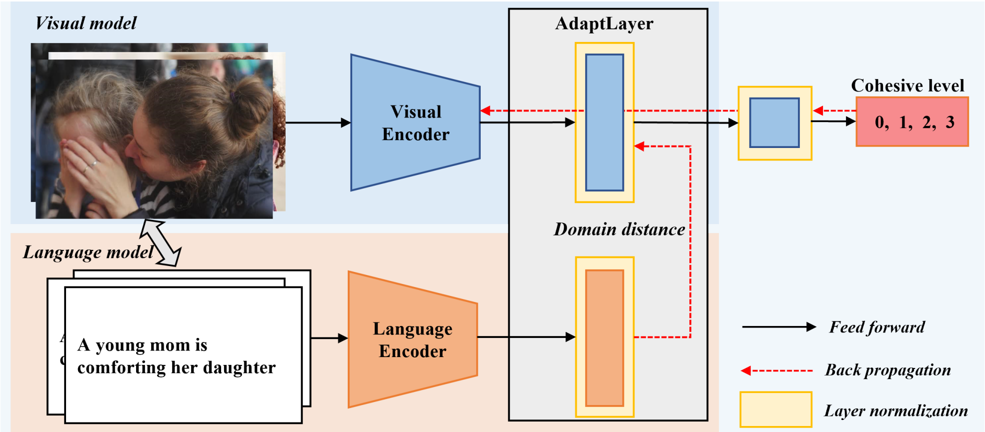


Figure 5.4: A knowledgeLayer ablation network for ablative Study 1.

Visual Encoder	visual LDNN (AL)		visual LDNN (AL+KL)	
	Val	Test	Val	Test
ResNet50	**0.6884**	**0.5295**	**0.6892**	**0.4677**
ResNet152	0.7279	0.5402	0.7032	0.4678
DenseNet161	0.7375	0.5623	0.6896	0.4820

Table 5.6: Comparison results of ablative study 1. “AL” and “KL” denote the AdaptLayer and KnowledgeLayer, respectively.

Ablative Study 2: We further analyze the impact of the language model by comparing an LDNN composed of two visual encoders (LDNN[V + V]) with a regular LDNN[V + L] using one visual encoder and one language encoder. As the results in Tab. 5.7 show, all the visual models trained on LDNN [V + L] show much lower MSE than those trained on LDNN [V + V]. These findings confirm that the linguistic knowledge of cohesive levels trained by the language model could affect the performance of the visual model. As was made clear by ablative study 1, we believe that the linguistic knowledge in descriptions was injected into the visual model through domain adaptation and linguistic knowledge transition mechanisms. In addition, we compared the AlexNet-based and ResNet50-based visual LDNNs with a typical multimodal model (Fusion) [40] that combines visual and language features in the final regression layer. The results of the validation set provided in Tab. 5.7 show that visual LDNNs could achieve comparable to those obtained using the fusion model, even though the visual LDNN does not use language features for inference. We can accordingly affirm that LDNN is an effective method to inject linguistic knowledge into visual models.

Visual Encoder	LDNN (V+L)		LDNN (V+V)		Fusion
	Val	Test	Val	Test	Val
AlexNet	0.7432	0.5167	0.8016	0.6125	0.7481
VGG11	0.7556	0.5099	0.7747	0.5394	-
ResNet18	**0.6774**	0.4893	0.7747	**0.5394**	-
ResNet50	0.6892	**0.4677**	**0.7112**	0.5485	**0.6835**

Table 5.7: Comparison results of ablative study 2.

Effective size: Ablation studies have proved that LDNN is an effective way to inject linguistic knowledge of language models into visual models. On the other hand, how much linguistic knowledge could be injected into the visual model also needs to be clarified. We compared visual LDNNs that are trained with the same language encoder and various visual encoders, and show the comparison results for the test set in Fig. 5.5. We define the MSE gap between the visual LDNN and fine-tuned DNN as effective size. We noticed that the effective size is different from the visual encoders, and the visual LDNN trained with a high MSE visual encoder has a large effective size. This could be explained by the fact that those high MSE visual encoders are not sufficient to obtain meaningful representations so it is easy to inject the linguistic knowledge into the visual model. In contrast, we also consider that the language model performance is another limitation on effective size. Since the

Japanese-based language encoder has an MSE (shown in Tab. 5.3) comparable to that of the visual encoder (shown in Tab. 5.4), we believe that the current language model is insufficient to provide much more meaningful linguistic knowledge to affect the visual model training.

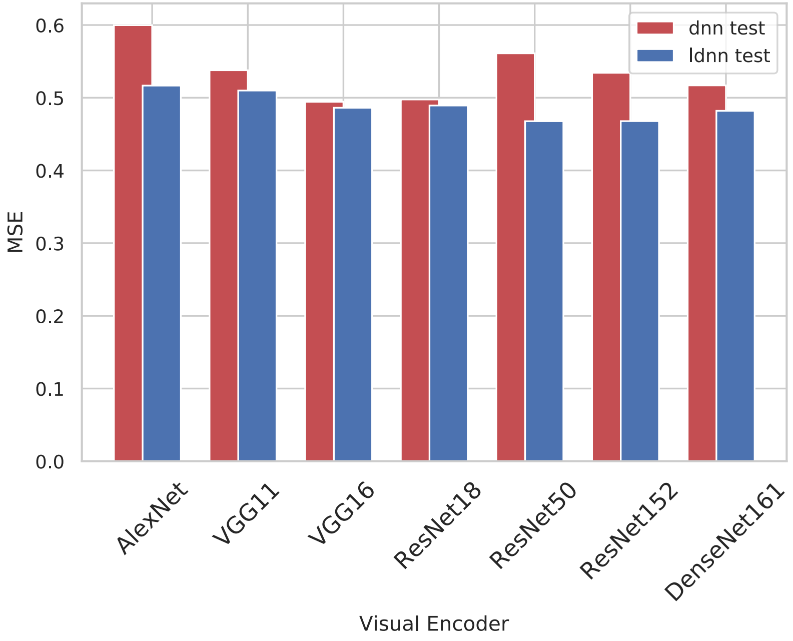


Figure 5.5: Comparison results of LDNN and DNN baselines on the testing set.

5.5 Conclusion

We proposed a linguistic knowledge injectable deep neural network (LDNN) that can build a visual model for predicting group cohesiveness that could automatically associate related linguistic knowledge hidden behind images. LDNN consisted of a visual encoder and a language encoder and applied domain adaptation and linguistic knowledge transition mechanisms to transform linguistic knowledge by training visual LDNN and a language model together. We evaluated LDNN on the GAF 3.0 dataset, and the results show that the visual model not only improves the performance of the fine-tuned DNN model leading to an MSE very similar to the state-of-the-art model but is also a practical and efficient method that requires relatively little preprocessing. Furthermore, ablation studies confirmed that LDNN is an effective method for injecting linguistic knowledge into visual models.

Chapter 6

Implicit Knowledge Injectable Cross Attention Audiovisual Model for Group Emotion Recognition

In this chapter, we extend LDNN by additionally injecting audio and language knowledge encoded by pretrained models into a multimodal model. Here, we present a knowledge injection audiovisual network that facilitates knowledge distillation across diverse modalities, leading to comprehensive multimodal knowledge transformation.

6.1 Introduction

Group emotion recognition is a critical step towards future artificial intelligence (AI) technologies that can understand complex human relationships and enable high-level interaction with humans [149]. Unlike Inferring individual emotions in videos by applying visual face region pixels and audio frequency domain features (Melspectrogram, MFCCs, etc.) [59, 40, 150, 151], group emotions have been inferred by applying multiple visual features such as faces, scenes, skeletons, and objects [152, 153]. However, group emotion is a reflection of human interactions that involve intentions, activities, and relationships within the group, and those explicit visual features are insufficient to represent complex human interactions.

According to the mechanism by which the conceptual knowledge accumulated in the brain is transformed into a visual cognition process [154, 155, 156, 157], humans can unconsciously understand the context of a video beyond the audiovisual information presented in the video. Therefore, we believe that various potential information related to human interactions can be used to gain meaningful knowledge to facilitate an understanding of group emotions. For example, in Fig. 6.1, the emotion being expressed in the video can be correctly recognized as positive. Not only can we see objects such as three men, one woman, and an L-shaped sofa, but we can also understand the context that "two men and a woman sit on an L-shaped sofa and talk to each other on a TV show".

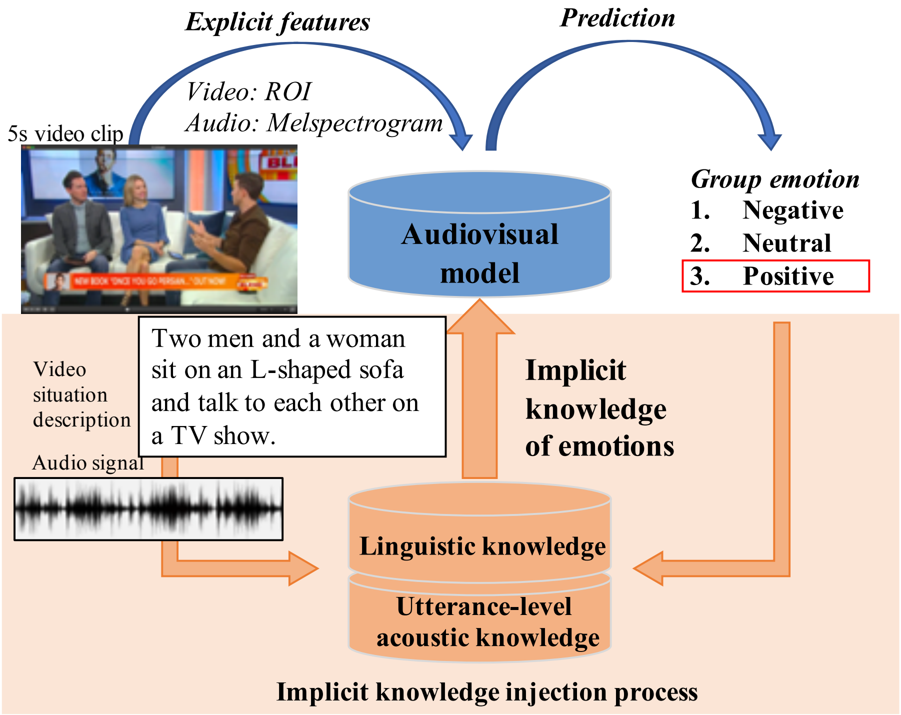

Explicit features
Prediction
Video: ROI
Audio: Melspectrogram
5s video clip
Audiovisual model
Group emotion
1. Negative
2. Neutral
3. Positive
Video situation description
Two men and a woman sit on an L-shaped sofa and talk to each other on a TV show.
Implicit knowledge of emotions
Audio signal
Linguistic knowledge
Utterance-level acoustic knowledge
Implicit knowledge injection process

Figure 6.1: A conceptual figure demonstrating the process of injecting implicit knowledge into an audiovisual model, and the emotion prediction by using only explicit features.

In this chapter, we define the information that is presented in the video as explicit information, and the features extracted from them as explicit features. We apply two kinds of explicit features: 1) The ROI features represent the region of the objects that are present in the video, and these features are extracted using a pretrained model directly [142]; 2) Melspectrogram features are extracted by processing the raw audio signals contained in the video [158]. In contrast, we define the linguistic and acoustic emotional representations that do not exist in the audio-video data as implicit knowledge. Human language contains a rich source of knowledge that can be used to describe human interactions from the perspectives of intentions, activities, and relationships within the group. We convert the video situation descriptions to word embeddings using the BERT pretraining model [72] and distill emotional linguistic knowledge from them. In addition, i-vector is a kind of utterance-level acoustic feature that can explain the variability of channel, speaker, language, and emotion [159]. We construct and process i-vector features with linear discriminant analysis (LDA) [160] from basic acoustic features (MFCCs, pitch, and energy) and further distill emotional acoustic knowledge from them.

As implicit knowledge does not exist in the video and cannot be used to directly infer group emotions, we propose an end-to-end architecture, called the implicit knowledge injectable cross-attention audiovisual deep neural network (K-injection audiovisual network) to train an audiovisual model that can not only obtain audiovisual representations of group emotions through an explicit feature-based cross attention audiovisual subnetwork (audiovisual subnetwork) but also can absorb implicit knowledge of emotions through two implicit knowledge-based injection subnetworks (K-injection subnetwork). As a result, though the K-injection audiovisual network is trained with explicit features and implicit knowledge, it is easy to apply since it does not require any implicit knowledge during inference.

As shown in Fig. 6.1, the K-injection audiovisual network is built by integrating the audiovisual subnetwork and two K-injection subnetworks. The audiovisual subnetwork is a multi-head cross-attention network, which takes explicit features as the input and dynamically integrates both explicit features throughout the sequence. On the other hand, the K-injection subnetwork distills implicit linguistic and acoustic knowledge corresponding to the emotion target from video situation descriptions and basic acoustic features and aims to transform them into the audiovisual model. We trained the K-injection audiovisual network on the training set and achieved an overall accuracy of 66.19% for the validation outperforming the baseline accuracy for the validation set of 50.05% [125]. Furthermore, we take the average of the audiovisual models trained with the (linguistic, acoustic, and linguistic-acoustic) K-injection subnetworks for the testing set, and the overall accuracy is 66.40% compared to the baseline accuracy of 47.88%.

Our contribution to this chapter is summarized as follows: 1) We propose an end-to-end architecture that can not only obtain audiovisual representations from the video directly but also can absorb implicit knowledge of emotions hidden in the video; 2) We apply a multi-head cross-attention subnetwork as the audiovisual subnetwork that can dynamically integrate multimodal features throughout the sequence level; 3) We apply two knowledge-based injection subnetworks that can transform the knowledge distilled from an unimodal model and transfer it into another model; 4) We train an audiovisual model that is a simple model that only requires explicit features during inference.

6.2 Related works

Most recent studies on emotion recognition have aimed to assign emotion labels to one person in a video [59, 161, 40]. Although it is becoming recognized as common to build an audiovisual emotion model using visual face region pixels and audio frequency domain features, how to effectively fuse these audio-video features for obtaining expressive multimodal representations is still a critical issue. These works [58, 40] employ attention mechanisms for integrating word-level multimodal representations. In particular, the work [59] proposes a multiple attention-based network to dynamically select the key representations of each modality throughout the sequence for multimodal fusion. In contrast, our model focuses on how to additionally transfer multimodal knowledge encoded by pretrained unimodal encoders into a multi-head cross-attention network to further enhance the performance.

Preparing various features from diverse modalities is crucial for obtaining expressive representations. For achieving group-level cohesion prediction from the visual scene, these works [152, 153, 129, 130, 131] have tried several visual features such as faces, scenes, skeletons, and objects [142]. As group cohesion is the reflection of human interaction, explicit visual features alone are not sufficient for obtaining holistic representations, and the background knowledge hidden in the image, especially the implicit knowledge related to human interactions, can be expected to be a key factor. Instead of extracting linguistic knowledge by diverting image captioning tasks that are not suitable for acquiring linguistic knowledge of group emotion [144, 145], we propose two K-injection subnetworks to learn implicit linguistic knowledge and utterance-level acoustic knowledge representing group emotion from the video description and utterance-level audio signals. This allows us to train an audiovisual model that can incorporate this implicit knowledge.

Recent knowledge distillation approaches have been proposed for training lightweight models by transferring knowledge from a large model into a smaller model [147], and the cross-modal transfer was designed for training a speech emotion recognition model by distilling the knowledge of a pretrained facial emotion recognition network [162]. These studies demonstrate that the knowledge extracted from pretrained models can be transferred or injected into other models, regardless of whether the model applies the same modality. Our proposed K-injection subnetworks are built by distilling the implicit knowledge representing group emotions and transferring it into a multimodal audiovisual model. Thereby, once the audiovisual model has been trained, no implicit knowledge is required during inference.

6.3 Method

To learn an audiovisual model that can not only obtain meaningful representations from explicit features (i.e., visual and audio features) but also obtain pretraind implicit K-injection linguistic and acoustic knowledge representing group emotions, our method contains an audiovisual subnetwork and two knowledge-injection subnetworks (a linguistic K-injection subnetwork and an acoustic K-injection subnetwork). We explain the detail of our proposed K-injection audiovisual network as shown in Fig. 6.2.

6.3.1 Audiovisual subnetwork

Explicit features: We build a video feature extractor following [142, 163] in taking the features of detected objects as video features $feat_v^o$. A 5-second video is first clipped into n frames, and each frame f_v is further detected m objects $\{o_1, ..., o_m\}$. Each object o_i is finally converted to the ROI pooling features with 2048 dimensions. Here, we use a frame size of 40 and set the maximum number of objects at 18, the video features of a video are shaped as $(40, 18, 2048)$; In terms of the audio feature extraction, we build an audio feature extractor that utilizes ffmpeg command to convert the 5-second video into the waveform audio data with a sampling rate of 44.1KHz, and then extract 128-dimensional melspectrogram features from every single frame f_a as the audio features $feat_a^f$. Since the audio frame size is set to 0.02 seconds and the shift size is set to 0.01 seconds, we finally get a melspectrogram feature sequence with a shape of $(501, 128)$.

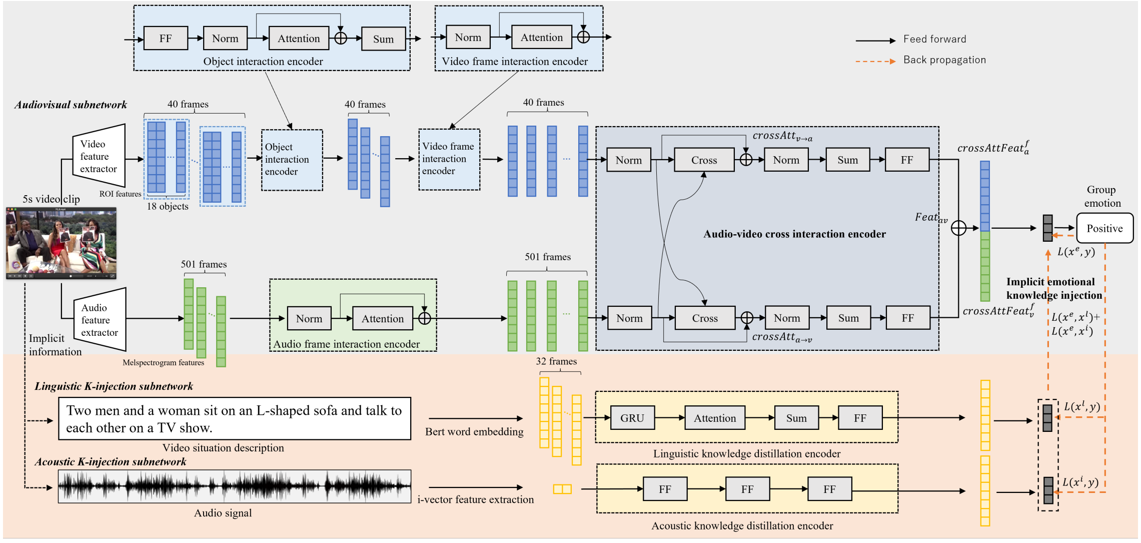


Figure 6.2: K-injection audiovisual network: An end-to-end architecture for training an audiovisual model in which the implicit knowledge representing the emotions of a group is transferred from two K-injection subnetworks. The top part shows the audiovisual subnetwork. "FF", "Norm", and "Sum" denote the feedforward layer, normalization method, and attention-weighted sum calculator, respectively. "Attention" indicates multi-head self attention and "Cross" indicates multi-head cross attention. The bottom part shows the linguistic K-injection subnetwork and acoustic K-injection subnetwork. "GRU" denotes the gated recurrent unit layer, and "Self" indicates self-attention. The symbol of "⊕" presents a residual connection.

Architecture: The audiovisual subnetwork is composed of the object/video-frame/audio-frame interaction encoders and an audio-video cross-interaction encoder (shown in Fig. 6.2). All the encoders are built on a multi-head attention mechanism [21], which consists of two parts, the scaled dot-product attention and multi-head attention. As Eq. 6.1 shows, the scaled dot-product attention is computed by packing a batch of queries, keys, and values into matrices Q, K, and V, respectively. It takes these matrices as the inputs and outputs weighted values, where the weights are computed by applying a softmax function. In Eq. 6.2, the multi-head attention is computed to jointly capture multiple correlations between the query and key.

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}(QK^T)V \tag{6.1}$$

$$\begin{aligned} &MultiHead(Q, K, V) = \mathrm{Concat}(head_1, ..., head_h)W^o, \\ &where\ head_i = \mathrm{Att}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{6.2}$$

where d_k and d_v are the dimension size of keys and values matrices; The parameter matrices are $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and

$W_i^o \in \mathbb{R}^{hd_v \times d_{model}}$. In this study, we set $d_{model} = 128$ and $h = 2$, and thus $d_k = d_v = d_{model}/h = 64$.

The object/video-frame/audio-frame interaction encoders are built by the multi-head self-attention layer, and all of the keys, values, and queries are the same as each other. The object interaction encoder is intended to dynamically highlight objects based on the interaction of all objects within each frame. All the Q, K, V is $feat_v^o$, and the output of the object interaction encoder is computed as a weighted sum of $feat_v^o$ throughout all objects, called $feat_v^f$ with a shape of $(40, 2048)$. The video/audio frame interaction encoders are also built by the multi-head self-attention layer and assign the frame sequence feature $feat_v^f$, $feat_a^f$ to the attention layer, respectively. The video/audio frame attentional features $featAtt_v^f$, $featAtt_a^f$ are computed by assigning attention weights to each frame of $feat_v^f$ and $feat_a^f$.

Different from the multi-head self-attention, we build an audio-video cross-interaction encoder, a bidirectional multi-head cross-attention architecture, to dynamically obtain audiovisual representations corresponding to the target emotion. Following the equations Eq. 6.3 and 6.4, we compute $crossAtt_{v\to a}$ by inputting $featAtt_v^f$ as Q, $featAtt_a^f$ as K and V, and compute $crossAtt_{a\to v}$ by inputting $featAtt_a^f$ as Q, $featAtt_v^f$ as K and V, where both $crossAtt$ are followed by the normalization layer to avoid large domain distances. Finally, the audiovisual representations $feat_{av}$ are computed by concatenating the output of the feedforward layers which is applied to the weighted sum of $crossAtt_{v\to a}$ and $crossAtt_{a\to v}$, named as $crossAttFeat_a^f$ and $crossAttFeat_v^f$, respectively.

$$crossAtt_{v\to a} = \mathrm{MultiHead}(featAtt_v^f, featAtt_a^f, featAtt_a^f) \tag{6.3}$$

$$crossAtt_{a\to v} = \mathrm{MultiHead}(featAtt_a^f, featAtt_v^f, featAtt_v^f) \tag{6.4}$$

$$feat_{av} = \mathrm{Concat}(crossAttFeat_a^f, crossAttFeat_v^f) \tag{6.5}$$

6.3.2 K-injection subnetwork

Implicit knowledge: We defined the linguistic and acoustic implicit knowledge that does not exist in the video but is related to the video context in § 6.1. We add video situation descriptions[1] from the viewpoint of human interactions, and process these descriptions to obtain word embedding features using the BERT pretraining

[1] https://github.com/wangyanan326/Additional-EmotiW-dataset

model [72]. And then we apply these embedding features to the linguistic K-injection subnetwork to distill emotional linguistic knowledge. Since the video descriptions are created in Japanese, we use the BERT model provided by Kyoto University which is pretrained on the Japanese Wikipedia corpus. As a result, we can extract BERT embedding features with shape (L, D) from each piece of video data, where L denotes the number of words in the description, and D denotes the dimension of the word embedding vector. Here, we upsample the words to a maximum description length of 16 words, each word is a 768-dimensional vector. We construct and process i-vector features with LDA using video level group affect (VGAF) [127] to further improve the discrimination ability across emotion classes. The i-vector features yield a 2-dimensional feature vector and adapt to the acoustic K-injection subnetwork to distill emotional acoustic knowledge.

Architecture: The K-injection subnetwork is built by a knowledge distillation encoder with a knowledge injection loss function (shown at the bottom of Fig. 6.2). The linguistic knowledge distillation encoder takes BERT embedding features as input and follows “GRU”, “Self”, “Sum”, and “FF” layers to distill linguistic emotional knowledge from the emotion target. On the other hand, the acoustic knowledge distillation encoder applies i-vector acoustic features to a 3-layer neural network to distill acoustic emotional knowledge. The outputs of both the linguistic and acoustic knowledge distillation encoders are further calculated by a feedforward layer with three output units and finally passed to the softmax function. We call the last layer of the K-injection subnetwork the knowledge injection layer, where the distilled knowledge can be transferred to the audiovisual subnetwork through a knowledge injection loss function $\mathcal{L}^k$. In this chapter, we adopt the negative log-likelihood loss as the knowledge injection loss, and the linguistic and acoustic knowledge injection loss is defined as:

$$\ell(x, y) = -\log(\mathrm{softmax}(x_y)) \tag{6.6}$$

$$\mathcal{L}^k = \ell(x^e, y^l) + \ell(x^e, y^a) \tag{6.7}$$

where the x^e is the output of the last layer of the audiovisual subnetwork, the y^l and y^a denote the prediction results of the linguistic and acoustic K-injection subnetworks, respectively. Both y^l and y^a are in the range $[0, C-1]$, where C is the number of classes (positive, neutral, negative).

6.3.3 Training and Inference

As described above, the K-injection audiovisual network is constructed by the audiovisual subnetwork and two K-injection subnetworks, and thus it is a multi-task learning network. We minimize the following loss $\mathcal{L}$ at the backpropagation process to train these subnetworks jointly. Here, we use the negative log-likelihood loss as the emotion classification loss as well as the knowledge injection loss.

$$\mathcal{L} = \ell(x^e, y) + \ell(x^l, y) + \ell(x^a, y) + \mathcal{L}^k \tag{6.8}$$

where x^l and x^a denote the output of the linguistic and acoustic K-injection subnetwork, respectively, and y is the index of the target emotion class in the range $[0, 3]$.

We train an audiovisual model that can not only obtain meaningful representations from explicit features but also acquire emotional knowledge from implicit knowledge. In the inference, we only utilize the audiovisual model with the input of explicit information to predict the group's emotions.

6.4 Experiment

In this section, we introduce the task, dataset, baseline, and experimental settings, and then evaluate the performance of our audiovisual K-injection subnetwork on the VGAF validation dataset [127]. In addition, we explain the ensemble strategies for obtaining robust results on the testing set.

6.4.1 Task, dataset

The Audio-video Group Emotion Recognition task is to assign a single emotion label from three classes (positive, neutral, and negative) to each video [127]. The video data are collected from YouTube and have been clipped into 5-second lengths, referred to as the Video level Group AFfect (VGAF) dataset [127]. In the VGAF dataset, each video contains a group of people who are singing, talking, protesting, fighting, or engaged in other activities (A sample video frame is shown in Tab .6.2). The VGAF dataset shown in Tab. 6.1 consists of Training (2,661), Validation (766), and Testing sets (756).

Dataset	Total	Positive	Neutral	Negative
Train	2,661	802	923	936
Valid	766	302	280	184
Test	756	-	-	-

Table 6.1: The size of training (Train), validation (Valid), and testing (Test) sets.

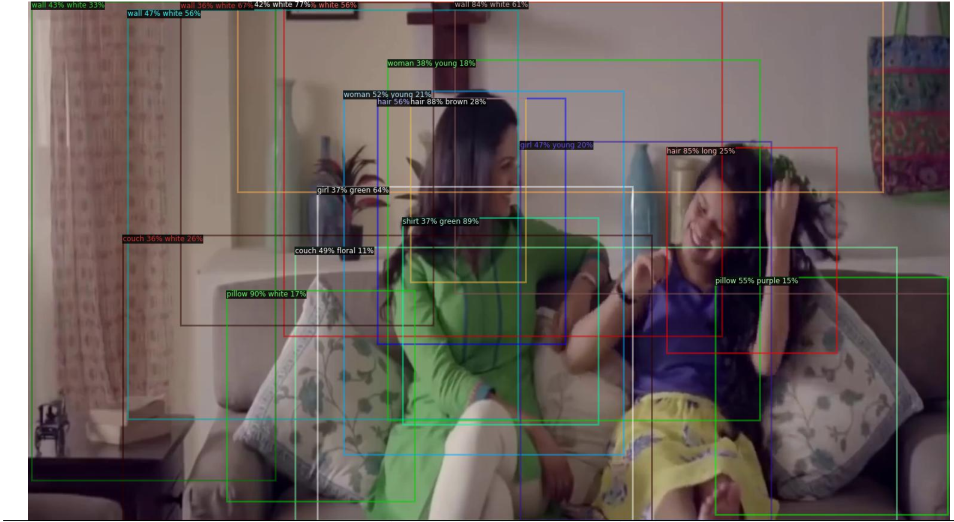
Video frame
 wall 43% white 33% wall 36% white 67% 42% white 77% % white 56% wall 84% white 61% wall 47% white 56% woman 38% young 18% woman 52% young 21% hair 56% hair 88% brown 28% girl 47% young 20% hair 85% long 25% girl 37% green 64% shirt 37% green 89% couch 36% white 26% couch 49% floral 11% pillow 90% white 17% pillow 55% purple 15%
Video caption: 青い服の女性が、緑の服の女性に甘えるようにしている (A woman in green is amusing a woman in blue)
Label: Positive

Table 6.2: A video frame randomly selected from the VGAF validation set, the bounding boxes on the image indicate the region of objects. We added the video situation description in the Japanese language and used it as implicit information for training.

6.4.2 Baseline

The baseline model is an inception-based network that was trained on an image-based group-level emotion dataset (GAF) [125]. Video frame features are extracted using the pretrained inception model, and audio features are extracted based on the ComParE challenge feature set [164]. The baseline model was built as an LSTM and fully connected based network taking the video frame features and audio features as inputs, respectively. The classification accuracy of the baseline for the validation set is 50.05%.

6.4.3 Experimental settings

To optimally train our method, as shown in Fig. 6.2, we adopt a normalization layer for the object/video-frame/audio-frame interaction encoders and the audio-video cross-interaction encoder to avoid large domain distances, and also process residual connections in these encoders to integrate more information. For the multi-head attention mechanism, we experimentally set the number of heads as 2 and the dimension of input as 128. A dropout rate of 0.5 is applied to all the encoders and the K-injection subnetworks. Since our method is a multiple learning network, we set an optimal learning rate (LR) for the audiovisual subnetwork and linguistic/acoustic K-injection subnetwork, respectively. The LR for the audiovisual subnetwork is 0.00001, the linguistic K-injection subnetwork is 0.01, and the acoustic K-injection subnetwork is 0.001. In addition, we train our method for 200 epochs on the training set and also train each fold for 150 epochs based on the cross-validation training (details in §6.4.4).

6.4.4 Ensemble strategy

We adopt 7-fold, 8-fold, and 9-fold cross-validation methods to generate various prediction results for the testing set [165]. Here, the stratified cross-validation used to keep the class distribution is the same in each fold. The average of the k-fold prediction results for the testing set is computed for subsequent ensemble processing. Since we train 3 kinds of audiovisual models, i.e., the linguistic K-injection model, acoustic K-injection model, and linguistic-acoustic K-injection model, the average of these models for the testing set is taken as the first submission. Furthermore, the second submission is the 7-fold cross-validation model, the third submission is the 8-fold cross-validation model and the fourth submission is the 9-fold cross-validation model. We finally take the average of the first/second/third/fourth submissions as the fifth submission.

6.4.5 Results on the validation set

Unimodal models: We trained various unimodal models with inputs of explicit and implicit modal features, and show the accuracy for the validation set in Tab. 6.3. The visual modal models took ROI features (bounding boxes shown in Tab. 6.2), Resnet152 features [17] and DenseNet161 features [146] as inputs, and were trained based on the same Bi-GRU attention architecture. It is particularly worthy that the ROI features have a higher score than the others. The reason for this outcome can be attributed to the fact that the ROI features are pretrained bottom-up attention features which contain far more reasonable features than pretrained CNN features[142]. From the results of audio modal models, it is obvious that the implicit acoustic features (i-vector) have much higher accuracy than the explicit audio features (Melspectrogram). In addition, the results based on the video situation description (BERT embedding) demonstrate that implicit linguistic information can lead to much higher accuracy compared to video-audio information. That is why we adopt this implicit information to distill emotional knowledge. We also annotated ground truth for the validation set and compared it with the actual labels, As shown in Tab. 6.3, the human score overwhelmed all unimodal models. It is a challenging task and no unimodal models can exceed human intelligence.

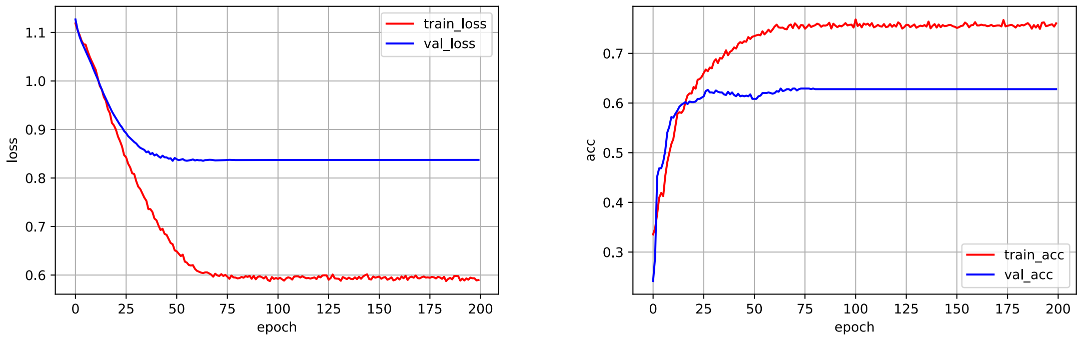
Modality	Model	Feature	Val(Acc)
Vision	Dual Bi-GRU attention	ROI	0.6084
	Bi-GRU attention	Resnet152	0.5614
	Bi-GRU attention	DenseNet161	0.6070
Audio	Bi-GRU attention	Melspectrogram	0.5183
	3-layer neural network	i-vector	0.5561
Language	Self-Attention	BERT embedding	0.6501

Table 6.3: Comparison results of unimodal models on the validation set.

Modality	Model	Feature		Validation (Acc)		
		explicit	implicit	BERT embedding	i-vector	audiovisual
human score	-	-	-	-	-	0.7802
vision, audio	LSTM-based network (Baseline)	Inception visual feature, ComParE audio feature	-	-	-	0.5005
vision, audio	cross-multi head attention	ROI, Mel spectrogram	-	-	-	0.6279
vision, audio	K-injection network	ROI, Mel spectrogram	i-vector	-	0.5509	0.6358
vision, audio, lang	K-injection network	ROI, Mel spectrogram	BERT embedding	0.5640	-	0.6527
vision, audio, lang	K-injection network	ROI, Mel spectrogram	i-vector, BERT embedding	0.5470	0.5525	**0.6619**

Table 6.4: Comparison results with a baseline on the validation set.

Multimodal models: We trained our model by applying the above unimodal features (explicit and implicit features), and show the results in Tab. 6.4 for the accuracy of audiovisual models where the inputs were solely explicit features. Compared to the cross-multi head attention model (the top part of the Fig. 6.2) trained with explicit features (ROI and Mel spectrogram), all the K-injection subnetwork-based audiovisual model have much higher accuracy for the validation set. In particular, the K-injection subnetwork trained with both implicit features (i-vector and BERT embedding) achieved the highest accuracy of 66.19% compared to other multimodal models, and this is 16.14% higher than the baseline model. In addition, we compare the training progress of the cross-multihead attention network (Fig. 6.3) and K-injection subnetworks (Fig. 6.4). From the training loss progress, the K-injection subnetwork-based audiovisual model is hard to train in the initial epochs. This is because our method is a multiple learning network and the learning speed of each subnetwork is difficult to control. Once the training of all subnetworks has converged, the loss value of the K-injection subnetwork becomes lower than that of the cross-multihead attention network. The K-injection subnetwork with both implicit features has the lowest loss value, and its accuracy is higher than unimodal models. Therefore, we believe that our proposed K-injection subnetwork is effective in achieving an audiovisual model that is superior to all unimodal and multimodal models by integrating implicit knowledge during inference.



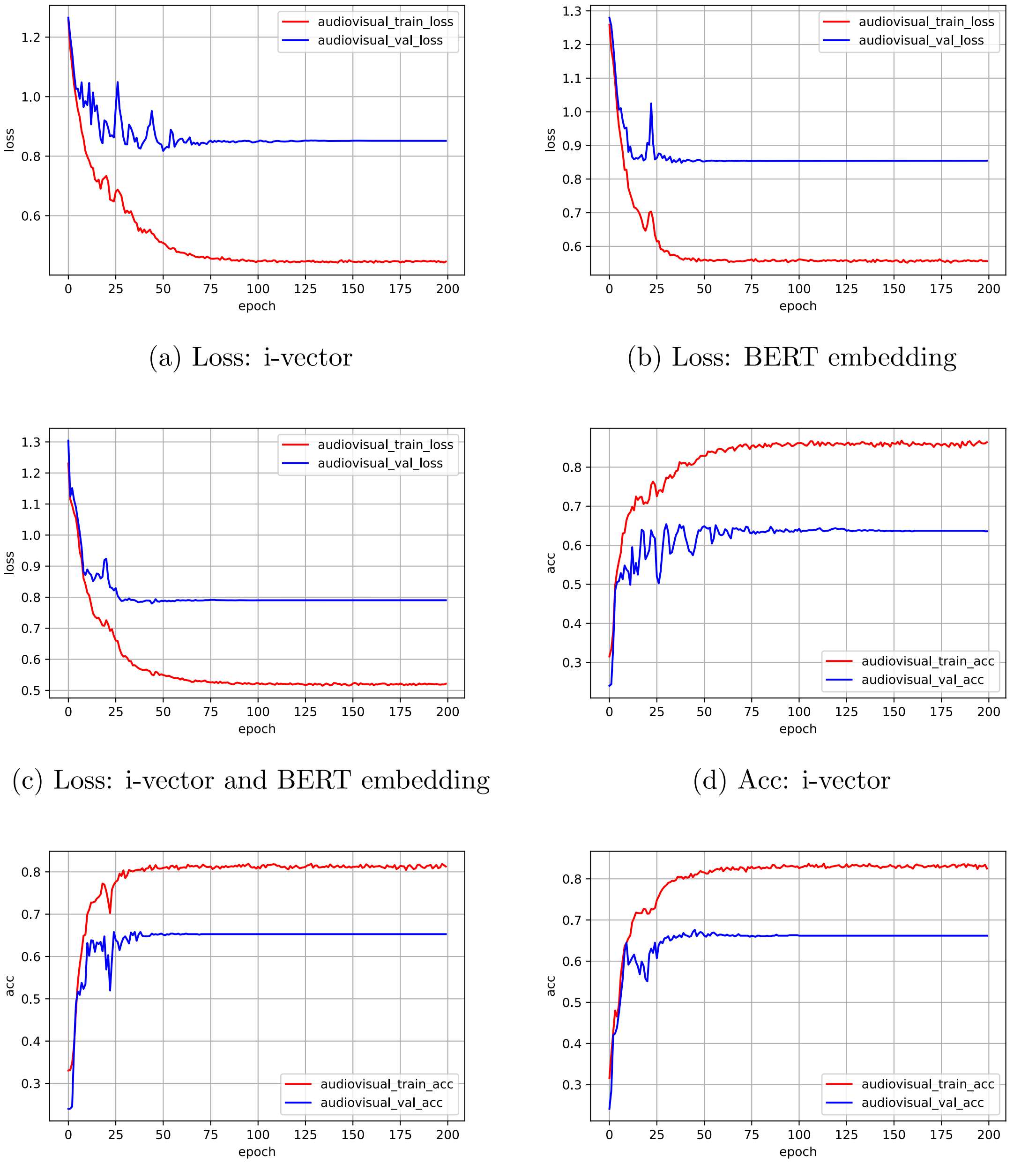
(a) Loss: i-vector

(b) Loss: BERT embedding

Figure 6.3: Training progress of the cross-multi head attention based audiovisual model

6.4.6 Ensemble results on the testing set

Following the ensemble strategies (section 6.4.4), we obtained the accuracy results of all submissions. As shown in Tab. 6.5, submission 1 achieved the highest score of 66.40%, which is the average of the prediction results of the three K-injection



(a) Loss: i-vector

(b) Loss: BERT embedding

(c) Loss: i-vector and BERT embedding

(d) Acc: i-vector

(e) Acc: BERT embedding

(f) Acc: i-vector and BERT embedding

Figure 6.4: Training progress of the K-injection subnetwork-based audiovisual model.

subnetwork-based audiovisual models. On the other hand, the other submissions using the cross-validation method achieved much higher classwise accuracy on the neutral and negative classes, but much lower positive class accuracy than submission 1. As a consequence, the overall accuracy of the cross-validation method did not achieve the expected results. We noticed that some of the video data clipped from the same video existed in different folds so the cross-validation methods were not able to learn more discriminative emotional representations. We should have separated the validation set under the premise of no video clips coming from the same video data to further improve the performance.

Submission	Classwise (Test: acc)			Overall
	Positive	Neutral	Negative	(Test: acc)
1	**0.5576**	0.7864	0.6000	**0.6640**
2	0.4470	0.7994	0.6000	0.6376
3	0.4793	0.7735	**0.6174**	0.6415
4	0.4148	**0.8188**	0.6000	0.6362
5	0.4700	0.8058	0.6087	0.6495

Table 6.5: Submission accuracy for the testing set.

6.5 Conclusion

We proposed a K-injection audiovisual network to train an audiovisual model that can not only obtain audiovisual representations of group emotions through the audiovisual subnetwork but is also able to absorb implicit knowledge of emotions through two K-injection subnetworks. The K-injection audiovisual network was trained by applying explicit features (ROI and Melspectrogram) and implicit features (BERT embedding and i-vectors) as the input. It is an efficient method that only requires explicit features during inference.

Chapter 7

VideoAdviser: Video Knowledge Distillation for Multimodal Transfer Learning

In chapter 5 and chapter 6, two knowledge injection approaches were discussed. These approaches involve distilling pretrained unimodal knowledge into other modal models to create efficient multimodal models. However, the success of these approaches is constrained by the quality of the pretrained knowledge and the effectiveness of the chosen multimodal knowledge distillation strategy.

In this chapter, to address the above issues to achieve high efficiency-performance multimodal transfer learning, we propose VideoAdviser, a video knowledge distillation method to transfer multimodal knowledge of video-enhanced prompts from a multimodal fundamental model (teacher) to a specific modal fundamental model (student). With an intuition that the best learning performance comes with professional advisers and smart students, we use a CLIP-based teacher model to provide expressive multimodal knowledge supervision signals to a RoBERTa-based student model via optimizing a step-distillation objective loss—first step: the teacher distills multimodal knowledge of video-enhanced prompts from classification logits to a regression logit—second step: the multimodal knowledge is distilled from the regression logit of the teacher to the student. We evaluate our method in two challenging multimodal tasks: video-level sentiment analysis (MOSI and MOSEI datasets) and audio-visual retrieval (VEGAS dataset). The student (requiring only the text modality as input) achieves an MAE score improvement of up to **12.3%** for MOSI and MOSEI. Our method further enhances the state-of-the-art method by **3.4%** mAP score for VEGAS without additional computations for inference.

7.1 Introduction

Transfer learning is a promising methodology that focuses on transferring pretrained representation domains to nearby target domains [166]. For instance, finetuning a pretrained language model on a small annotated dataset enables high-performance text sentiment analysis [167]. Recent fundamental models on diverse modalities such as language models (*e.g.*, RoBERTa [5], GPT-3 [22]), visual models (*e.g.*, ViT [4]), and multimodal models (*e.g.*, CLIP [29], MEET [168]) have millions of parameters and can provide robust modal representations. With such advancement, multimodal transfer learning aims to transform pretrained representations of diverse modalities into a common domain space for effective multimodal fusion [169, 162]. It has been broadly applied to multimodal tasks such as video-level sentiment analysis [63, 170, 171], and audio/text-video retrieval tasks [172, 173, 174, 175].

Existing works on multimodal transfer learning unify adversarial learning to regularize the embedding distributions between different modalities, leading to effective multimodal fusion [176, 177, 178, 173, 1]. However, conventional systems are typically built on the assumption that all modalities exist, and the lack of modalities always leads to poor inference performance. For instance, vision-language models typically fail to achieve expected performance when given only text data as input. Furthermore, extracting pretrained embeddings for all modalities is computationally inefficient for inference. Therefore, improving robust multimodal transfer learning to achieve high efficiency-performance inference is crucial for practical applications, which motivates this work.

Knowledge distillation (KD) is first proposed for achieving an efficient student model by transforming embedded knowledge in the predicted logits of the teacher model to a smaller student model [147]. Recent works have expanded it to multimodal transfer learning by distilling mutual information from one modality to another [179, 180, 181, 182, 183]. However, these works always need to sacrifice the performance of the teacher model, requiring the teacher model and the student model distributed in neighboring domains (*e.g.*, vision→vison, text→text).

In this paper, with an intuition that the best learning performance comes with professional advisers and smart students, to achieve high efficiency-performance multimodal knowledge distillation, we propose VideoAdviser shown in Figure 7.1, a

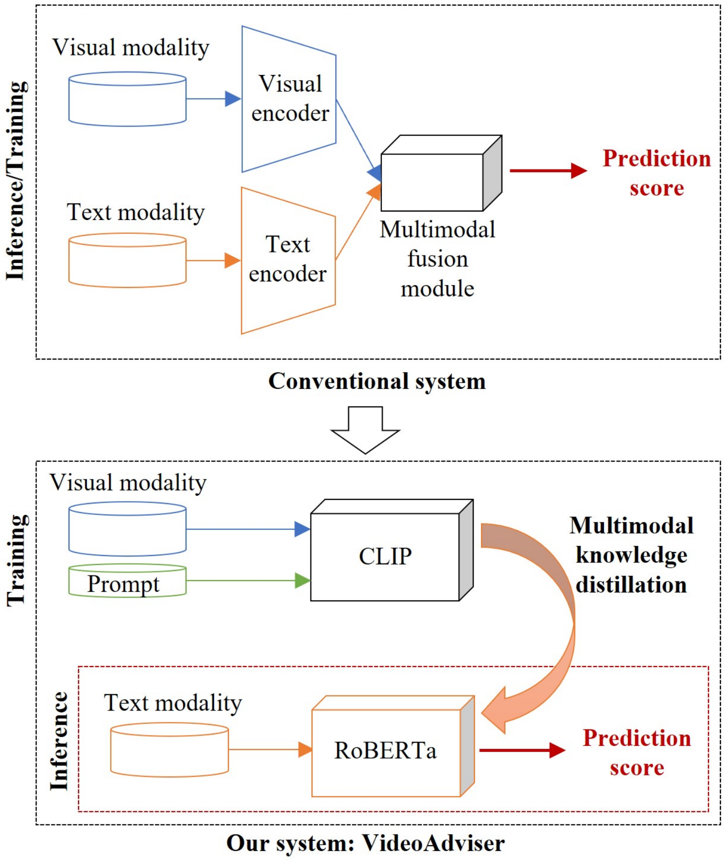


Figure 7.1: A conceptual diagram illustrates the difference between the conventional system and our system: our system focuses on transferring multimodal knowledge from a multimodal fundamental model (*e.g.*, CLIP) to a language fundamental model (*e.g.*, RoBERTa-Large), and requires text only to achieve high efficiency-performance inference. On the other hand, the conventional system focuses on multimodal fusion and requires complex modules (diverse modal encoders and a multimodal fusion module) for inference.

video knowledge distillation method to transfer multimodal knowledge from a strong multimodal fundamental model (teacher) to a powerful specific modal fundamental model (student) via optimizing a step-distillation objective loss. As CLIP is a multimodal fundamental model pretrained with cross-modal contrastive learning on tremendous image-text pairs [29], we employ it as the teacher model to obtain multimodal knowledge of video-enhanced prompts by incorporating the video and text prompt representations. The teacher model utilizes CLIP's visual and text encoders to obtain video and text prompt embeddings without freezing the pretrained weights to preserve multimodal representation space learned by CLIP. By adapting transformer-based modules on these embeddings and extracted frame-level facial expression features, the teacher model acquires expressive multimodal knowledge of video-enhanced prompts by performing video and text prompt representations learning. To sufficiently absorb distilled multimodal knowledge from the teacher model, we employ a large-scale language model RoBERTa [5] as the student model. Since RoBERTa is a transformer-based architecture composed of huge parameters, we finetune its full parameters to leverage RoBERTa's powerful architecture to achieve high-performance student models for inference. In addition, we propose a step-distillation objective loss to distill coarse-fine grained multimodal knowledge to further improve the multimodal knowledge distillation. Motivated by multiscale representation learning enabling the fusion of enriched coarse-fine grained representations [184, 185], we consider that multitask with different target granularities allows the model to acquire representative knowledge at diverse granularities. For instance, classification encourages the model to separate the data point into multiple categorical classes representing an interval of consecutive real values to acquire knowledge at a coarse granularity. In contrast, regression enables the model to distinguish the data point into continuous real values instead of using classes to learn knowledge at a fine granularity. To this end, in the first step, the teacher model distills multimodal knowledge of video-enhanced prompts from classification logits to a regression logit to unify knowledge at both coarse and fine granularity; In the second step, the unified multimodal knowledge is further distilled from the teacher model to the student model.

We evaluate VideoAdviser in two challenging multimodal tasks: video-level sentiment analysis (MOSI and MOSEI datasets) and audio-visual retrieval (VEGAS dataset). The RoBERTa-based student model requiring only text data as input outperforms the state-of-the-art multimodal model's MAE score by **12.3%** for MOSI and **2.4%** for MOSEI. Our method also enhances the state-of-the-art audio-visual cross-modal model by **3.4%** mAP score for VEGAS without additional computations for inference. Ablation studies further demonstrate that our method is able to improve the state-of-the-art method's MAE score by over **3.0%** with almost half the parameters. These results suggest the strengths of our method for achieving high efficiency-performance multimodal transfer learning.

7.2 Related work

Multimodal fundamental model

CLIP [29] is a multimodal fundamental model that learns transferable visual models from natural language supervision on a dataset of 400 million (image, text) pairs. It jointly trains an image encoder and a text encoder using contrasting learning objectives to obtain a joint multimodal representation space. Inspired by its remarkable zero-shot generation ability for downstream image tasks, the work [186] proposes XCLIP to expand pretrained CLIP on general video recognition by finetuning it on video data using a video-specific prompting module that enhances the video representation to the text representation. The work [187] utilizes a pretrained CLIP for open-vocabulary object detection by distilling visual knowledge from cropped image regions. In this work, we adapt a pretrained CLIP on distilling multimodal knowledge of video-enhanced prompts from the teacher model to the student model via a step-distillation objective loss.

Knowledge distillation based transfer Learning

In addition to achieving a lightweight student model by minimizing the KL divergence between the probabilistic outputs of a teacher and student model [147], recent works on knowledge distillation focus on transferring representational knowledge from a teacher model to a student model [188, 187, 189]. For instance, these works [190, 191] distill linguistic knowledge from a text encoder to a visual encoder by learning the mapping between modal representations. The work [8] utilizes multiple text encoders to perform cross-modal knowledge distillation for stronger text-video retrieval. The work [192] distills expressive text representations from a generation model to the text encoder of CLIP by minimizing text-text feature distance. However, these works mostly focus on knowledge distillation in the common modal domain or show limited performance in the cross-modal domain. In contrast, to achieve expressive knowledge distillation for multimodal transfer learning tasks, we propose a RoBERTa-based student model to improve multimodal knowledge distillation by leveraging its powerful transformer architecture.

Video-level sentiment analysis task

Recent works [167, 170, 63] on video-level sentiment analysis tasks focus on improving modality fusion. The work [177] proposes VAE-Based adversarial learning method to map multimodal representations to a joint domain space for improving the modality fusion process. The work [171] achieves SOTA performance on MOSI [2] and MOSEI [3] dataset by introducing a pretrained modality fusion module that fuses multimodal representation from multi-level textual information by injecting acoustic and visual signals into a text encoder. However, all these works require preprocessed multimodal embeddings as the input which is inefficient for inference. In contrast, we employ a knowledge distillation approach that requires only one specific modality leading to efficient inference.

Audio-visual retrieval task

Recent works on audio-visual retrieval tasks exploit supervised representation learning methods to generate new features across modalities in a common space [174, 193, 194, 195, 173, 172, 175, 11], such that the audio-visual features can be measured directly. Inspired by the C-CCA [193] that aims at finding linear transformations for each modality, C-DCCA [194] tries to learn non-linear features in the common space by using deep learning methods. Deep learning methods by using rank loss to optimize the predicted distances, such as TNN-C-CCA [172], and CCTL [175] models, which apply triplet losses as the objective functions to achieve better results than other CCA-variant methods. The EICS model [11] learns two different common spaces to capture modality-common and modality-specific features, which achieves the SOTA results so far. In this paper, we enable our method to enhance the extracted audio and visual representations of the SOTA model by distilling multimodal knowledge from a CLIP-based teacher model.

7.3 Problem setting

This work focuses on video-level sentiment analysis and audio-visual retrieval tasks, respectively. For the video-level sentiment analysis task, each data point consists of a video $\boldsymbol{M}$, the cropped sequential face images I, the divided speech text T_{speech}, and the class text T_{class}, our goal is to predict the sentiment intensity $\mathcal{Z}_{pred} \in [-3, 3]$ by giving only speech text T_{speech} for inference. For the audio-visual retrieval task, assume that $\Gamma = \{\gamma_i\}_{i=1}^{N}$ is a video collection, $\gamma_i = \{a_i, v_i\}$, where N indicates the data size, $a_i \in \mathbb{R}^{D1}$ and $v_i \in \mathbb{R}^{D2}$ are audio and visual features from different feature spaces. Our target aims at feeding them into a common space by mapping functions $f(x)$ and $g(x)$ to generate new features $f(a_i)$ and $g(v_i)$. As a result, each query a_i for example will obtain a rank list from another modality based on *query*-$v_j(i \neq j)$ similarity.

7.4 Methodology

In this section, we explain our method VideoAdviser in detail. As shown in Fig. 7.2, our method consists of a CLIP-based model as the teacher (§ 7.4.1) and a RoBERTa-based model as the student (§ 7.4.2). The teacher and student models are jointly trained to achieve knowledge distillation across modalities. The student model enables sentiment intensity prediction by giving only a speech text for inference (§ 7.4.3). We use $\mathcal{F}(\cdot)$, $\mathcal{V}(\cdot)$, $\mathcal{P}(\cdot)$ and $\mathcal{T}(\cdot)$ to denote the facial expression encoder, visual encoder, prompt encoder, and text encoder.

7.4.1 The CLIP-based teacher model

Facial expression embedding To enhance the visual representations of the teacher model for sentiment intensity prediction, we first use OpenFace [77] to crop face images $\{I_i\}_{i=1}^{T} \in \mathbb{R}^{P^2 \times 3}$ with each of size $P \times P$ pixels from T sampled video frames, then, we extract frame-level facial expression embedding $\boldsymbol{v}^{(f)} \in \mathbb{R}^{T \times D}$ with a facial expression encoder $\mathcal{F}(\cdot)$ [78] that is pretrained on the VGG-Face dataset [79]. Here, $\boldsymbol{v}^{(f)}$ is an 8-dimensional sequential vector of length 64 $[T = 64,\ D = 8]$. More details of the pretrained model on Albanie's website [1].

$$\boldsymbol{v}^{(f)} = \mathcal{F}(\{I_i\}_{i=1}^{T}) \tag{7.1}$$

Visual embedding To fully transfer the powerful generality of pretrained CLIP [29] from image to video, we freeze the parameters of pretrained CLIP visual encoder $\mathcal{V}(\cdot)$ to obtain frame-level visual embedding $\boldsymbol{v}^{(v)} \in \mathbb{R}^{T \times D}$, where T denotes the number of sampled video frames and D is the dimension of visual embedding. Following [186], given a video clip $M \in \mathbb{R}^{T \times H \times W \times 3}$ of T sampled video frames with $H \times W$ pixels, we use ViT-L/14 [4] to first divide t-th frame into N patches $\{x_{t,i}\}_{i=1}^{N} \in \mathbb{R}^{P^2 \times 3}$, where $t \in T$ and $N = HW/P^2$. Then, the patches $\{x_{t,i}\}_{i=1}^{N}$ is mapped to $\boldsymbol{v}^{(v)} = \{v_t^{(v)}\}_{t=1}^{T}$

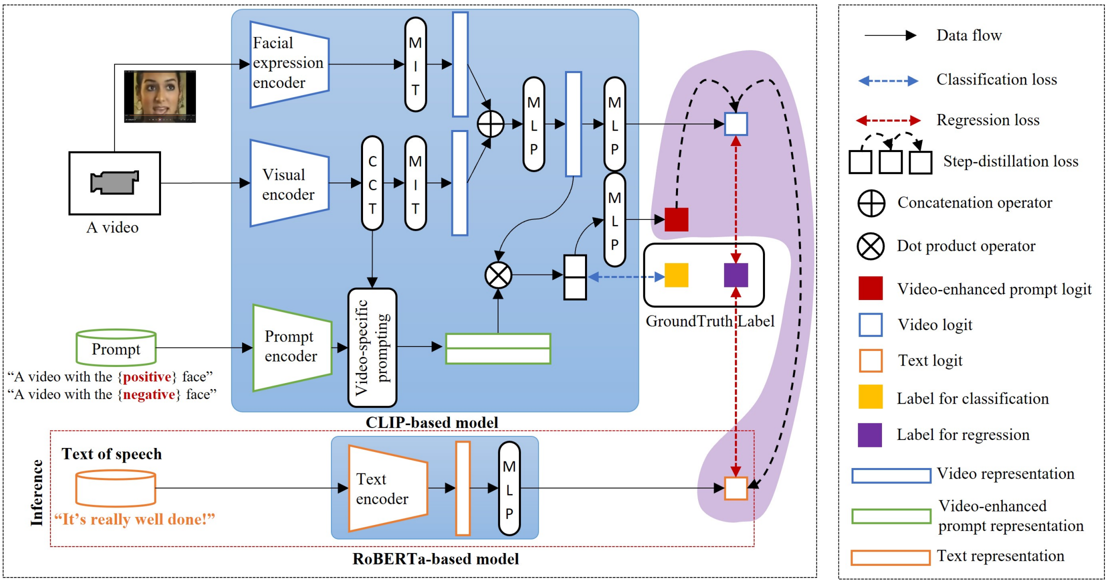


Figure 7.2: Architecture of VideoAdviser using a CLIP-based model (the teacher) to distill multimodal knowledge of video-enhanced prompts to a RoBERTa-based model (the student): the teacher model utilizes pretrained CLIP's text and visual encoders, and a facial expression encoder to obtain the sentiment class text embedding, the frame-level embedding, and the facial expression embedding. Then, the teacher model employs CCT, MIT, MLP, and a video-specific prompting module, and minimizes a binary sentiment classification loss and a sentiment regression loss. Meanwhile, the student model is finetuned on speech text by minimizing a sentiment regression loss and a step-distillation loss (the region in purple). During inference, the speech text is used to enable sentiment intensity prediction. Here, CCT, MIT, and MLP stand for the cross-frame communication transformer, multi-frame integration transformer, and multi-layer perceptron, respectively.

with a linear transformation $f_m : \mathbb{R}^{P^2 \times 3} \to \mathbb{R}^{3P^2 \times D}$.

$$\boldsymbol{v}^{(v)} = \mathcal{V}(f_m(\{\boldsymbol{x}_t\}_{t=1}^{T})) \tag{7.2}$$

Text prompt embedding We employ the text encoder $\mathcal{P}(\cdot)$ of pretrained CLIP to obtain text prompt embedding $\boldsymbol{v}^{(p)} \in \mathbb{R}^{C \times D}$ of C sentiment classes by giving the sentiment class label $T_{class} \in \{\text{negative,positive}\}$, where "positive" class includes 0. The text prompt such as "A video with the $\{T_{class}\}$ face" is generated with a text prompt generator f_g and encoded as

$$\boldsymbol{v}^{(p)} = \mathcal{P}(f_g(T_{class})) \tag{7.3}$$

We employ the cross-frame communication transformer (CCT), multi-frame integration transformer (MIT), and video-specific prompting modules to obtain expressive multimodal sentiment knowledge. The CCT is a multi-layer transformer with cross-frame attention introduced in [186] to enable cross-frame information exchange. It is used to obtain cross-frame visual representations by giving a modified visual embedding $\bar{\boldsymbol{v}}^{(v)} = \{\bar{\boldsymbol{v}}_t^{(v)}\}_{t=1}^{T}$, where $\bar{\boldsymbol{v}}_t^{(v)} = [x_{class}, v_t^{(v)}] + \boldsymbol{e}_{pos}$. x_{class} is a learnable frame representation and e_{pos} is a position embedding of patches in a frame. The MIT is a normal transformer layer constructed by standard multi-head self-attention and feed-forward networks. Given frame-level embeddings $\boldsymbol{v}^{(f)}$ and $\bar{\boldsymbol{v}}^{(v)}$, we finally obtain the video representation V as follows:

$$V^{(f)} = \text{AvgPool}(\text{MIT}(\boldsymbol{v}^{(f)})) \tag{7.4}$$

$$V^{(v)} = \text{AvgPool}(\text{MIT}(\text{CCT}(\bar{\boldsymbol{v}}^{(v)}))) \tag{7.5}$$

$$V = f_v([V^{(f)}||V^{(v)}]) \tag{7.6}$$

where $f_v : \mathbb{R}^{2\mathcal{D}} \rightarrow \mathbb{R}^{\mathcal{D}}$ is a two-layer MLP. AvgPool denotes an average pooling layer. "||" denotes a concatenation operator used to process facial expression-conditioned video representation. We then transform the **video representation** V to the **video logit** (see Fig. 7.2) with a two-layer MLP.

Inspired by [186], the teacher model employs a video-specific prompting module to enhance the prompt embedding with cross-frame visual representations. The video-specific prompting module applies a normal multi-head attention [21] to obtain the **video-enhanced prompt representation** $\bar{\boldsymbol{v}}^{(p)} \in \mathbb{R}^{C \times D}$ (see Fig. 7.2) as

$$\bar{\boldsymbol{v}}^{(p)} = \boldsymbol{v}^{(p)} + \text{Multi\_Head\_Attention}(\text{CCT}(\bar{v}^{(v)})) \tag{7.7}$$

Then, we compute dot product between video representation V and video-specific prompt representation $\bar{\boldsymbol{v}}^{(p)} = \{\bar{\boldsymbol{v}}_i^{(p)}\}_{i=1}^{C}$ to output the similarity score $\boldsymbol{p} = \{p_i\}_{i=1}^{C}$ with a softmax layer as

$$p_i = \text{softmax}(\bar{v}_i^{(p)} \cdot V) = \frac{\exp(\bar{v}_i^{(p)} \cdot V)}{\sum_{i \in C} \exp(\bar{v}_i^{(p)} \cdot V)} \tag{7.8}$$

where C indicates the number of sentiment classes. We further transform $\boldsymbol{p}$ into the **video-enhanced prompt logit** (see Fig. 7.2) with a two-layer MLP.

7.4.2 The RoBERTa-based student model

To leverage the powerful transformer-based architecture of fundamental language models, we structure a RoBERTa-based student model [5] that consists of a text encoder $\mathcal{T}(\cdot)$ and a two-layer MLP. Given the speech text T_{speech}, the student model obtains text representation $V^{(t)}$ with $\mathcal{T}(\cdot)$, and output sentiment intensity $\mathcal{Z}_{pred}$ with the MLP into the **text logit** (see Fig. 7.2) as

$$\mathcal{Z}_{pred} = \text{logit}(V^{(t)}), V^{(t)} = \mathcal{T}(T_{speech}) \tag{7.9}$$

Where $V^{(t)} \in \mathbb{R}^{D}$, and $\text{logit}(\cdot) : \mathbb{R}^{\mathcal{D}} \to \mathbb{R}^{1}$ indicates the two-layer MLP.

7.4.3 Training objectives

We simultaneously optimize the teacher and student models by applying mean squared error (MSE) loss to obtain video and text sentiment knowledge. Both teacher and student models minimize the L_2 distance as follows:

$$\mathcal{L}_v^{(r)} = \text{MSE}\left(\text{logit}(V), l^{(r)}\right), \ \mathcal{L}_t^{(r)} = \text{MSE}\left(\mathcal{Z}_{pred}, l^{(r)}\right) \tag{7.10}$$

where $\mathcal{L}_v^{(r)}$ indicates MSE between the teacher model's video logit and sentiment label $l^{(r)}$, and $\mathcal{L}_t^{(r)}$ indicates MSE between the student model's text logit ($\mathcal{Z}_{pred}$) and $l^{(r)}$. Here, $\text{logit}(V)$ is a two-layer MLP for transforming video representation V into the video logit.

To learn the video-enhanced prompt representation to fuse multimodal knowledge of video and class text, we use the binary sentiment classification label $l^{(c)}$ (see Fig. 7.3) synthesized from the sentiment label to optimize the teacher model with a cross-entropy loss $\mathcal{L}_v^{(c)}$ as

$$\mathcal{L}_v^{(c)} = -\sum_{i=1}^{C} l_i^{(c)} \log(p_i), \tag{7.11}$$

We optimize a step-distillation objective loss to achieve multimodal knowledge distillation from the teacher model to the student model. The step-distillation objective loss consists of a **prompt-video distance minimization** $\mathcal{L}_{p\to v}$ and a **video-text distance minimization** $\mathcal{L}_{v\to t}$, where $\mathcal{L}_{p\to v}$ is optimized to align coarse-grained classification knowledge in the video-enhanced prompt logit and fine-grained regression knowledge in the video logit, $\mathcal{L}_{v \to t}$ is optimized to align knowledge in the video logit of the teacher model and the text logit of the student model. We apply MSE loss to perform the step-distillation as follows:

$$\mathcal{L}_{p \to v} = \mathrm{MSE}(\mathrm{logit}(\boldsymbol{p}), \mathrm{logit}(V)),\ \mathcal{L}_{v \to t} = \mathrm{MSE}(\mathrm{logit}(V), \mathcal{Z}_{pred}) \tag{7.12}$$

where logit($\boldsymbol{p}$) indicates the coarse-grained classification knowledge in Eq. 7.11.

We finally have a joint loss $\mathcal{L}$ for training the teacher and student models end-to-end as

$$\mathcal{L} = \alpha \mathcal{L}_v^{(r)} + \beta \mathcal{L}_t^{(r)} + \gamma \mathcal{L}_v^{(c)} + \delta \mathcal{L}_{p \to v} + \psi \mathcal{L}_{v \to t} \tag{7.13}$$

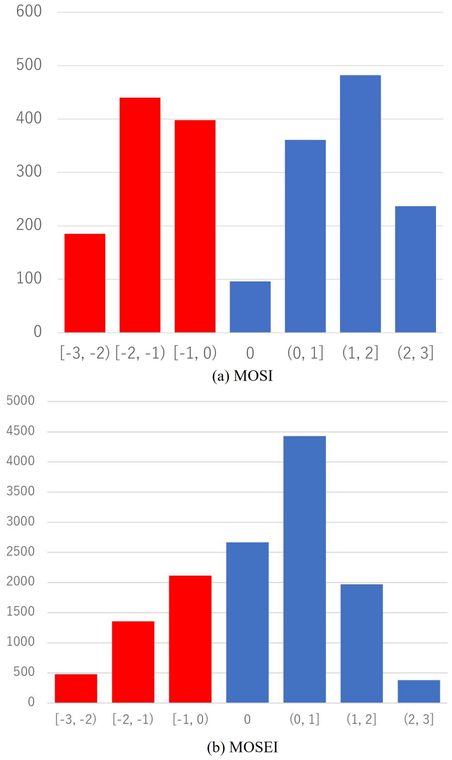
where α, β, γ, δ, and ψ indicate the importance of each loss value. They are empirically set as $1 : 10 : 1 : 10 : 1$ to keep all loss values on the same scale.

7.5 Experiment

In this section, we conducted empirical experiments on video-level sentiment analysis and audio-visual retrieval tasks to demonstrate the high efficiency-performance of our method.

7.5.1 Dataset

MOSI [2] and MOSEI [3] are multimodal datasets collected from online video for evaluating video-level sentiment analysis tasks. We show the dataset size in Tab. 7.1. MOSEI drops the data lacking modalities to fairly evaluate recent modality fusion-based methods [1]. We compared the video segment IDs of each data point for each modality and saved only the data points associated with a common segment ID. The modified MOSEI dataset was found to be more challenging than the original dataset as it lowered the strong baseline MSE score by 4.9% (see Tab. 7.3). Both datasets are annotated with a Likert scale in the range of $[-3, 3]$, *i.e.*, (-3: highly negative, -2: negative, -1: weakly negative, 0: neutral, +1: weakly positive, +2: positive, +3: highly positive). We further synthesize binary classification label, *i.e.*, ([-3,0): negative, [0,3]: non-negative) used for optimizing the teacher model (§7.4.1). The label distribution is illustrated in Fig. 7.3. MOSEI is imbalanced and over 65% of data is distributed in $[-1, 1]$.



(a) MOSI

(b) MOSEI

Figure 7.3: Label distribution of (a) MOSI and (b) MOSEI. The synthesized binary classification label is illustrated in different colors (the "negative" class in red color and the "non-negative" class in blue color).

VEGAS dataset [196] is applied for the audio-visual retrieval task, which contains 28,103 videos in total as shown in Tab. 7.1. Each video can be embedded as an audio feature vector and a visual feature vector, and the audio-visual pair shares the same single label. The label represents an audio event (*e.g.*, baby crying) of the human voice or natural sound. The number of label classes is 10, and the length of each audio-visual pair ranges from 2 to 10 seconds.

Dataset	Train	Validation	Test	Total
MOSI [2]	1,284	229	686	2,199
MOSEI [3]	9,473	1,206	2,710	13,389
VEGAS [196]	22,482	-	5,621	28,103

Table 7.1: Dataset size. MOSEI uses the same dataset as [1].

7.5.2 Evaluation metric

We use the mean absolute error (MAE), accuracy (A^7), accuracy (A^2), and weight-$F1$ score for evaluating MOSI and MOSEI. A^7 denotes a 7-class and A^2 denotes a binary accuracy metric. Since MOSI and MOSEI are regression problems, we consider MAE to be the most reasonable metric for fair evaluations. In addition to the binary accuracy reported by most of the previous works, we evaluate the 7-class accuracy as did the SOTA method [171] to eliminate the effect of the data imbalance. For the audio-visual retrieval task, we apply the mean average precision (mAP) as previous works [11, 175] to evaluate our model.

7.5.3 Training setting

We train the teacher and the student models simultaneously and use only the student model for inference. The text modality is used for evaluating MOSI and MOSEI. On the other hand, as shown in Fig. 7.4, we utilize the teacher model to distill multimodal knowledge for both visual and audio encoders of the state-of-the-art model EICS [11] for audio-visual retrieval tasks. Both visual and audio encoders are used as student models to evaluate VEGAS. We show the hyperparameters of VideoAdviser (§7.4) for both tasks in detail in Tab. 7.2.

7.5.4 Performance

Evaluation of video-level sentiment analysis

We compared VideoAdviser with strong baseline methods on the test set of MOSI and MOSEI in Tab. 7.3. Compared with the state-of-the-art method UniMSE [171] that utilizes the powerful architecture of a large-scale pretraining model T5 [197] to improve the multimodal fusion by embedding multimodal signals into an auxiliary layer of T5, VideoAdviser is a multimodal knowledge distillation-based method that distills multimodal knowledge from a multimodal fundamental model CLIP [186] to a language model RoBERTa [5]. UniMSE was trained by integrating the training datasets of MOSI, MOSEI, MELD [198], IEMOCAP [199] and multimodal signals are required for inference. In contrast, our method was trained using the target dataset and requires only text data for inference. VideoAdviser significantly improves UniMSE's MAE score by **12.3%** for MOSI, and outperforms a strong baseline method VAE-AMDT's MAE score by **2.4%** for MOSEI. As we use the teacher

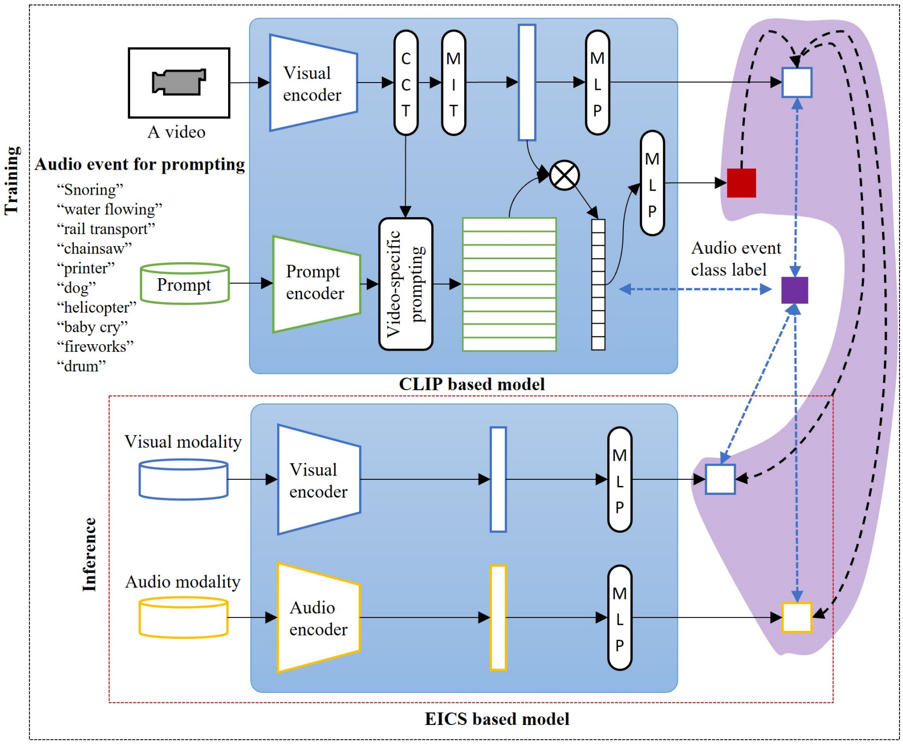


Figure 7.4: Architecture of VideoAdviser for audio-visual retrieval task using a CLIP-based model (the teacher) to distill multimodal knowledge of video-enhanced prompts to an EICS-based audio-visual model (the student). The teacher model is finetuned for the audio event classification to distill multimodal knowledge to the student model via the step-distillation loss (the region in purple). We adopt 3-layer MLP with 128-dimensional hidden layers.

model to offer auxiliary multimodal supervision signals to the student model, by leveraging the strengths of the learned multimodal space of the teacher model and the large-scale parameters of the student model, we think our method is effective for achieving high-performance multimodal knowledge distillation via minimizing the step-distillation objective loss (§7.4.3).

Evaluation of audio-visual retrieval

We further evaluated our VideoAdviser on the VEGAS dataset in Tab. 7.4. Compared to the state-of-the-art method EICS [11] that builds two different common spaces to learn the modality-common and modality-specific features, which achieves an average mAP of 0.788. Our method utilizes the distilled multimodal knowledge to enhance the performance of EICS. As a result, it achieves an average mAP of 0.822 and improves EICS [11] by **3.4%**, suggesting the generality of our method on audio-visual retrieval tasks.

	Hyperparameter	MOSI, MOSEI	VEGAS
Video	visual encoder	ViT-L/14	
	Num. of frames	8	
	Frame size	224×224	
	visual embedding size (input)	(B, 64, 8)	(B, 1, 10)
	Visual hidden layer size	(B, 128)	
Prompt	Prompt encoder	ClipTextModel	
	Prompt embedding size (input)	(B,77,512)	
	Prompt hidden layer size	128	
Text	Text encoder	RoBERTa-large	-
	Text embedding size (input)	(B,100,1024)	-
	Text hidden layer size	128	-
Audio	Audio encoder	-	EICS model
	Audio feature size (input)	-	10
	audio hidden layer size	-	128
Output logit	Video-enhanced prompt logit	(B, 1)	(B, 10)
	Video logit	(B, 1)	(B, 10)
	Text logit	(B, 1)	-
	Audio logit	-	(B, 10)
Optimizer	Method	AdamW [80]	
	Learning rate	8e-6	
	Warmup steps	15	
	Schedular	cosine_schedule_with_warmup	
Training	GPU	GTX 1080 Ti	
	Batch size	4	
	Training epochs	100	

Table 7.2: The hyperparameters for training VideoAdviser. Here, “B” denotes the batch size, “Audio logit” denotes the output of the audio encoder for VEGAS (see Fig. 7.4).

7.5.5 Efficiency

By comparing the number of parameters with state-of-the-art models in Tab. 7.5, our proposed VideoAdviser requires only a language model as the student that can achieve a high efficiency-performance model for inference. The Student (BERT [72]) achieved a compatible MAE score with fewer parameters than previous BERT-based models. Moreover, these models always process visual and audio signals for multimodal fusion, which might require more parameters and increase the computation cost. Compared with the state-of-the-art model UniMSE that uses a pre-trained transformer-based language model T5 [197] to perform multimodal fusion, our model, the student (ROBERTa-Base [5]) with nearly half of the parameters

Model	**MOSI**					**MOSEI**				
	MAE ↓	A^7 ↑	A^2 ↑	F1 ↑	Corr ↑	MAE ↓	A^7 ↑	A^2 ↑	F1 ↑	Corr ↑
MISA [167]	0.804	-	80.8	80.8	0.764	0.568	-	82.6	82.7	0.717
VAE-AMDT [1]	0.716	-	84.3	84.2	-	0.526*	-	82.8*	**87.5***	-
MAG-BERT [200]	0.712	-	84.2	84.1	0.796	0.539	-	84.7	84.5	-
Self-MM [63]	0.713	-	84.0	84.4	0.798	0.530/0.579*	-	82.8/84.6*	82.5/84.6*	0.765/-
MMM [170]	0.700	46.7	84.1	84.0	0.800	0.526	54.2	82.2	82.7	0.772
UniMSE [171]	0.691	48.7	85.9	85.3	0.809	0.523	54.4	**85.9**	85.8	0.773
VideoAdviser (ours)	**0.568**	**51.3**	**87.7**	**87.9**	**0.872**	**0.502***	**54.5***	84.5*	85.0*	**0.810***
Human	0.710	-	85.7	87.5	0.820	-	-	-	-	-

Table 7.3: Comparison results for MOSI and MOSEI. Our model reduces the state-of-the-art UniMSE's MAE score by **12.3%** for MOSI, and VAE-AMDT's MAE by **2.4%** for MOSEI. Here, (↓) indicates the lower the MAE, the better the performance, and (↑) indicates the vice-versa. (*) indicates the results produced on the modified MOSEI dataset.

reduces MAE score of over **3.0** point, suggesting the high efficiency-performance of our method. VideoAdviser was further improved over **9.0** point by adopting a RoBERTa-Large model as the student model.

7.5.6 Analysis

Effectiveness of components of the teacher model

We studied the effects of two core components of the teacher model (Facial expression encoder and video-specific prompting module) in Tab. 7.6. The results show that these two components help improve the multimodal knowledge distillation and boost the final performance of the student model. We believe that the facial expression encoder provided extra visual knowledge, and the video-specific prompting module further associated visual knowledge with text prompt representations encoded by the prompt encoder.

Effectiveness of the student model

We studied the effects of VideoAdviser on different student models in Tab. 7.7. We select two language models (BERT and RoBERTa) that have frequently been used in recent works [167, 1, 200, 63, 170]. By comparing the performance of language models with and without adopting a teacher model, the results demonstrate that our method improves a general language model's MAE score by over **6.0** point on average, suggesting the efficacy and generality of our method with different student models. We consider that the teacher model offers auxiliary multimodal supervision

Model	**VEGAS**		
	A→V	V→A	Average
Random	0.110	0.109	0.109
BiC-Net [174]	0.680	0.653	0.667
C-CCA [193]	0.711	0.704	0.708
C-DCCA [201]	0.722	0.716	0.719
DCIL [195]	0.726	0.722	0.724
DSCMR [173]	0.732	0.721	0.727
TNN-C-CCA [172]	0.751	0.738	0.745
CCTL [175]	0.766	0.765	0.766
EICS [11]	0.797	0.779	0.788
VideoAdviser (ours)	**0.825**	**0.819**	**0.822**

Table 7.4: The mAP comparison results with state-of-the-art models for VEGAS. Here, "V" and "A" indicate "Video" and "Audio", respectively.

to the student model during training, the language model-based students are able to learn multimodal knowledge from the teacher with their large-scale parameters.

We further trained a student model by freezing pretrained parameters, which dramatically dropped the MAE score from 0.568 to 1.478. This result makes us believe that in order to achieve expressive multimodal knowledge distillation across modalities, it is essential to finetune full parameters to leverage the strengths of large-scale pretrained models with powerful representational learning capabilities.

Modality effectiveness

To confirm the robustness of VideoAdviser in multimodal knowledge distillation not only for text modality but also for diverse modalities such as visual and audio modalities, we respectively studied the effects on visual and audio modalities for audio-visual retrieval tasks. As the results indicated in Tab. 7.8, the proposed step-distillation works for both modalities by boosting the baseline EICS model by over 1% mAP score. By associating both sides, we finally improved the baseline by 3.4%.

Model	Parameters	MOSI
		MAE
BERT-based model		
- MISA [167]	$> 110M$	0.804
- MAG-BERT [200]	$> 110M$	0.712
- Self-MM [63]	$> 110M$	0.713
- MMM [170]	$> 110M$	0.700
T5-based model		
- UniMSE [171]	$> 231M$	0.691
RoBERTa-based model		
- VAE-AMDT [1]	$> 355M$	0.716
VideoAdviser (ours)		
- Student (BERT)	$110M$	0.704
- Student (RoBERTa-Base)	$125M$	0.660
- Student (RoBERTa-Large)	$361M$	**0.568**

Table 7.5: Efficiency comparison. VideoAdviser is able to train a high efficiency-performance student model compared to state-of-the-art methods for inference. The student (RoBERTa-Base) outperforms the SOTA by over **3.0** point with nearly half the parameters.

Effectiveness of dataset size

In general, the larger the dataset, the better the performance. We trained VideoAdviser with a combination of the MOSI and MOSEI datasets to see if we can further improve the performance. As the results indicated in Tab. 7.9, The model performs much better than those trained on individual datasets and suggests the efficacy of our approach for different dataset sizes.

Effectiveness of the step-distillation loss

We ablatively studied the effects of our proposed step-distillation loss for multimodal knowledge distillation in Tab. 7.10. Without the first step—distilling multimodal knowledge from a video-enhanced prompt logit to a video logit (see Fig. 7.2), the learned multimodal space of CLIP cannot be passed to the student model via the video logit, resulting poor student model performance. On the other hand, it improves the regular language model (w/o step-distillation) **4.2%** MAE score and

Model	**MOSI**			
	MAE	A^7	A^2	F1
VideoAdviser (ours)	**0.568**	**51.3**	87.7	**87.9**
- w/o Facial expression encoder	0.579	50.2	86.8	86.4
- w/o Video-specific prompting	0.570	50.1	**88.1**	87.7

Table 7.6: Ablation results show the effects of components of the teacher model for multimodal knowledge distillation on MOSI dataset.

Model	**MOSI**		
	MAE	A^2	F1
Teacher (CLIP-based model)	-	57.3	-
BERT w/o teacher	0.753	84.1	83.6
Student (BERT)	0.704	84.7	83.8
RoBERTa-Base w/o teacher	0.719	84.6	84.3
Student (RoBERTa-Base)	0.660	85.4	84.6
RoBERTa-Large w/o teacher	0.660	87.3	87.3
Student (RoBERTa-Large)	**0.568**	**87.7**	**87.9**

Table 7.7: Effects in different student models. Our method improves the MAE score of pretrained language models by over **6.0** point on average.

suggests the effectiveness of the second step—distilling the knowledge of the video logit from the teacher model to the student model. Moreover, by optimizing the first and second steps, our proposed method outperforms a cutting-edge contrastive representation distillation method (CRD) [188] that proposed a contrastive-based objective for transferring knowledge between deep networks. Compared to the CRD which is designed to model mutual information across dimensions of the knowledge representations, Our proposed step-distillation applies MSE to mapping mutual information across modalities via one-dimensional logits (*i.e.*, video-enhanced prompt logit, video logit, and text logit). Our method performs better than CRD in transferring regression information for multimodal knowledge distillation.

In addition, we show comparison results of the proposed step-distillation loss with three widely-known distillation function KD [147], FitNet [202] and PKT [203] in Tab. 7.11. KD and PKT are proposed to minimize the KL divergence between the probabilistic outputs of a teacher and student model. On the other hand, FitNet

Model	**VEGAS**		
	A→V	V→A	Average
baseline (EICS [11])	0.797	0.779	0.788
VideoAdviser (ours)			
-w/ video distillation	0.794	0.810	0.802
-w/ audio distillation	0.791	0.815	0.803
-w/ (audio and video) distillation	**0.825**	**0.819**	**0.822**

Table 7.8: Ablation results show the effects of step-distillation on audio and video modalities for VEGAS. Here, "w/ video distillation" indicates that the step-distillation is only adopted for the visual modality of the student model, "w/ audio distillation" indicates the other side, and "w/ audio and video distillation" indicates both sides (see Fig. 7.4).

Test dataset	MAE	A^7	A^2
MOSI	**0.546** (0.568)	**51.3** (51.3)	**88.5** (87.7)
MOSEI	**0.491** (0.502)	**55.6** (54.5)	84.2 (**84.5**)
MOSI+MOSEI	0.502	54.79	85.05

Table 7.9: Results of VideoAdviser trained with a combination of MOSI and MOSEI datasets. The model performs much better for both the MOSI and MOSEI test sets. Here, (*) denotes the result of the model trained on the individual dataset.

and our step-distillation aim at minimizing the L_2 distance for knowledge distillation. Compared to KD, FitNet and PKT are one-step distillation loss functions, whereas our step-distillation performs two-step distillation, with the aim of transferring multimodal knowledge across multiple scales. To achieve a fair comparison, we adapted these three approaches to our problem setting of two-step distillation. As the results indicated in Tab. 7.11, the step-distillation outperforms other approaches and suggests its efficacy on multimodal knowledge distillation. We noted that the PKT-based two-step distillation achieves a compatible score with ours. We consider that audio-visual tasks focus on distilling multimodal knowledge of categorical audio events rather than fine-grained regressional knowledge so that transferring probabilistic knowledge of each category can also work well. Compared to KD which utilized the softmax function to obtain probabilistic knowledge, PKT adopted the cos-similarity function to better obtain dimension-level correlation with the probabilistic knowledge.

Model	**MOSI**			
	MAE	A^7	A^2	F1
CRD [188]	0.617	48.8	86.3	85.9
VideoAdviser (ours)				
- w/o step-distillation	0.660	45.5	87.3	87.3
- w/o step-distillation:step1	0.618	49.0	86.5	86.3
- w/ step-distillation	**0.568**	**51.3**	87.7	**87.9**

Table 7.10: Ablation results show the effects of the proposed step-distillation loss for MOSI.

Model	**VEGAS**		
	A→V	V→A	Average
KD [147]	0.783	0.612	0.701
FitNet [202]	0.803	0.781	0.792
PKT [203]	0.824	0.807	0.816
step-distillation (ours)	**0.825**	**0.819**	**0.822**

Table 7.11: Comparison results between widely-known knowledge distillation loss and the proposed step-distillation loss for VEGAS.

We further illustrate the logistic knowledge distribution with and without the step-distillation loss in Fig. 7.5. Compared to the "Text_logit w/o step-distillation" that plots the histogram of regression scores without performing the step-distillation, "Text_logit w/ step-distillation" is close to the groundTruth label distribution. Especially the distribution in the range of $[-1, 1]$ is strongly affected by the teacher model. Because the "Video_logit w/o step-distillation" distributes in the range of $[-1.5, 2]$ and the "Video_enhanced_prompt_logit w/o step-distillation" distributes in the range of $[-0.4, 0.2]$, by performing the step-distillation, the predicted regression score produced by the student model can be affected by the gap of these different distributions, and demonstrate that our proposed step-distillation is effective for multimodal knowledge distillation.

7.5.7 Significance Testing

We tested the stability of the performance improvement by VideoAdviser using the Almost Stochastic Order test (ASO) [204, 205] as implemented by [206]. We

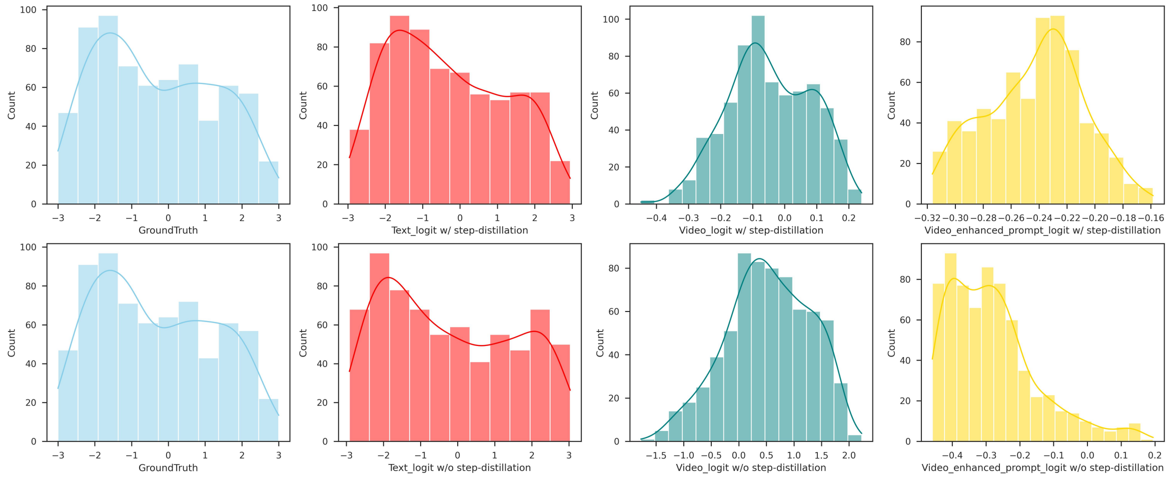


Figure 7.5: Visualization of logistic knowledge distribution with and without the step-distillation objective loss. The top row plots the histograms of logit by applying the step-distillation, and the bottom row indicates the vice-versa. The groudTruth indicates the label distribution, and text_logit indicates the predicted regression score of the student model. Our method using the step-distillation (the top) demonstrates a distribution of regression scores close to the groundTruth, affected by the knowledge distribution of the "video_logit" and "video_enhanced_prompt_logit".

compared three models, VideoAdviser (ours), VideoAdviser w/o step-distillation (baseline), and CRD based on five random seeds each using ASO with a confidence level of $\alpha = 0.05$. ASO computes a score (ϵ_{min}) indicated in Tab. 7.12 to represent how far the first model is from being significantly better with respect to the second. $\epsilon_{\text{min}} = 0$ represents truly stochastic dominance and $\epsilon_{\text{min}} < 0.5$ represents almost stochastic dominance.

Model	**ASO score (ϵ_{min})**
VideoAdviser (ours) → baseline	0
VideoAdviser (ours) → CRD	0
CRD → baseline	0.02

Table 7.12: ASO scores of models with different distillation objectives studied in Sec. 7.5.6. For "VideoAdviser (ours) → baseline", $\epsilon_{\text{min}} = 0$ indicates that VideoAdviser (ours) consistently outperform baseline. Here, the baseline denotes VideoAdviser (ours) w/o step-distillation.

7.6 Conclusion

We proposed a novel multimodal knowledge distillation method, VideoAdviser, which leverages the strengths of learned multimodal space of the CLIP-based teacher model and large-scale parameters of the RoBERTa-based student model to perform multimodal knowledge transfer by optimizing a step-distillation objective loss. In the evaluation of two multimodal tasks, our method significantly outperforms SoTA methods up to **12.3%** MAE score with a single modal encoder used in inference for video-level sentiment analysis, and **3.4%** mAP for audio-visual retrieval tasks, suggesting its strengths in high efficiency-performance. Ablation studies further demonstrate the efficacy of our proposed step-distillation objective loss in improving multimodal knowledge distillation. In the next step, we will adapt meta-learning to further explore the capability of multimodal transfer learning in a few-shot setting.

Chapter 8

Conclusion

This thesis covers the topics of high-performance multimodal fusion through the multimodal domain adaptation approaches (PART 1), and effective multimodal transfer learning to build efficient multimodal systems (PART 2).

In Chapter 3, we explained the limitations of current multimodal fusion approaches, and how the solution for transferring multimodality into a common domain space brings remedy to these limitations. In particular, we present the key contributions of VAE-AMDT: it performs VAE-based adversarial multimodal domain transfer learning to regularise devise modalities into a joint domain. This allows cross-modal attention architectures to compute expressive multimodal fusion. To achieve robust multimodal sentiment analysis systems, we plan to address some of the future challenges including the lack of commonsense knowledge related to the current video scene and inconsistent annotation for diverse modalities.

In Chapter 4, we propose a novel bidirectional fusion approach to enable conceptual reasoning of the model by effectively unifying a structured knowledge graph and unstructured pretrained language knowledge. It is built on a new multimodal GNN technique that performs inter-modal message passing to achieve expressive unified, multimodal graph representations. To further improve the temporal reasoning of current multimodal models, we plan to model hierarchical temporal relationships between objects in a video spatio-temporal graph.

We focus on building efficient multimodal systems in Part 2. In Chapter 5 and Chapter 6, we proposed two knowledge injection approaches based on knowledge distillation to enhance the performance of specific unimodal models. Moreover, these models demonstrate competitive performance with multimodal models on downstream tasks. To further address two issues: the quality of the pretrained knowledge and the effectiveness of the chosen multimodal knowledge distillation strategy, Chapter 7 shows that our proposed VideoAdviser handled these issues—leveraging the strengths of learned multimodal space of the CLIP-based teacher model and large-scale parameters of the RoBERTa-based student model—optimizing a step-distillation objective loss, achieved effective multimodal knowledge transfer learning. As one of our future works, we plan to unify meta learning to further enhance multimodal knowledge transfer learning.

Moreover, in the next step, seamlessly connecting multimodal knowledge aims to achieve fine-grained real-world understanding, we are planning to work on the following two topics: 1, unifying multimodal knowledge in structured graphs and fundamental models to enhance the interpretable reasoning of AI models. 2, building multimodal fundamental models that can not only solve general tasks but also can easily adapt to specific applications such as healthcare, autonomous vehicles, etc.

Bibliography

[1] Yanan Wang, Jianming Wu, Kazuaki Furumai, Shinya Wada, and Satoshi Kurihara. Vae-based adversarial multimodal domain transfer for video-level sentiment analysis. *IEEE Access*, 10:51315–51324, 2022.

[2] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[3] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2236–2246, 2018.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

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

[6] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.

[7] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836, 2021.

[8] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 11563–11573, 2021.

[9] A.-M. Oncescu, A.S. Koepke, J. Henriques, and Albanie S. Akata, Z. Audio retrieval with natural language queries. In *The annual conference of the International Speech Communication Association (Interspeech)*, 2021.

[10] A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685, 2023.

[11] Donghuo Zeng, Jianming Wu, Gen Hattori, Rong Xu, and Yi Yu. Learning explicit and implicit dual common subspaces for audio-visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(2), 2023.

[12] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[14] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2612–2620, 2017.

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[18] Jordi Pons and Xavier Serra. Timbre analysis of music audio signals with convolutional recurrent neural networks. *IEEE Transactions on Multimedia*, 19(4):822–833, 2017.

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

[20] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

[22] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.

[23] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems (NeurIPS)*, 29, 2016.

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[25] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*, 2020.

[26] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–137, 2020.

[27] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:23634–23651, 2021.

[28] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16375–16387, 2022.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763, 2021.

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900, 2022.

[31] Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: state of the art. *arXiv preprint arXiv:1612.02903*, 2016.

[32] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[33] Moataz M. H. El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44:572–587, 2011.

[34] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *The annual conference of the International Speech Communication Association (Interspeech)*, 2014.

[35] Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 115:24–35, 2018.

[36] Dhall Abhinav, Goecke Roland, Ghosh Shreya, and Gedeon Tom. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, 2019.

[37] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. Multi-feature based emotion recognition for video clips. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 630–634. ACM, 2018.

[38] Cheng Lu, Wenming Zheng, Chaolong Li, Chuangao Tang, Suyuan Liu, Simeng Yan, and Yuan Zong. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 646–652. ACM, 2018.

[39] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[40] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, E. Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential

learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[41] Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion. In *First Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language*, pages 116–125, 2018.

[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems (NeurIPS)*, 25, 2012.

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

[44] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[45] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017.

[46] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, and Jia Jia. Modality attention for end-to-end audio-visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6565–6569, 2019.

[47] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34–41, 2012.

[48] Albert Mehrabian. *Silent messages : implicit communication of emotions and attitudes.* Belmont, Calif. : Wadsworth Pub. Co., 1981.

[49] Yanan Wang, Jianming Wu, and Keiichiro Hoashi. Lightweight deep convolutional neural networks for facial expression recognition. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2019.

[50] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[51] Che-Wei Huang, Shrikanth Narayanan, et al. Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition. *arXiv preprint arXiv:1706.02901*, 2017.

[52] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[53] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details:delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[54] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing (ICONIP)*, pages 117–124, 2013.

[55] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 279–283, 2016.

[56] Philipp V. Rouast, Marc T. P. Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 12(2):524–543, 2021.

[57] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6558–6569, 2019.

[58] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7216–7223, 2019.

[59] Yanan Wang, Jianming Wu, and Keiichiro Hoashi. Multi-attention fusion network for video-based emotion recognition. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 595–601, 2019.

[60] Sijie Mai, Haifeng Hu, and Songlong Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 164–172, 2020.

[61] Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17(59):1–35, 2016.

[62] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[63] Wenmeng Yu, Hua Xu, Yuan Ziqi, and Wu Jiele. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[64] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[65] Youzhi Tu, Man wai Mak, and Jen-Tzung Chien. Variational domain adversarial learning for speaker verification. In *The annual conference of the International Speech Communication Association (Interspeech)*, 2019.

[66] Dae Hoe Kim, Wissam J. Baddar, Jinhyeok Jang, and Yong Man Ro. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, 10:223–236, 2019.

[67] Ming Li, Hao Xu, Xingchang Huang, Zhanmei Song, Xiaolin Liu, and Xin Li. Facial expression recognition with identity and emotion joint learning. *IEEE Transactions on Affective Computing*, 12:544–550, 2018.

[68] Zhengwei Huang, Ming Dong, Qi rong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 2014.

[69] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *International Conference on Platform Technology and Service (PlatCon)*, pages 1–5, 2017.

[70] Peng Song and Wenming Zheng. Feature selection based transfer subspace learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 11:373–382, 2020.

[71] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, 11:614–626, 2020.

[72] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019.

[73] Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[74] Jumayel Islam, Robert E. Mercer, and Lu Xiao. Multi-channel convolutional neural network for twitter emotion and sentiment recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[75] Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, Lili Mou, and Osmar R Zaiane. Seq2emo: A sequence to multi-label emotion classification model. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

[76] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems (NeurIPS)*, 27, 2014.

[77] Tadas Baltruaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

[78] S. Albanie and A. Vedaldi. Learning grimaces by watching tv. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[79] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[80] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2015.

[81] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

[82] Sinno Jialin Pan, Ivor Wai-Hung Tsang, James Tin-Yau Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22:199–210, 2009.

[83] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[84] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation

learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 104–120, 2020.

[85] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6720–6731, 2019.

[86] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[87] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14111–14121, 2021.

[88] Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1356–1365, 2021.

[89] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

[90] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3716–3725, 2020.

[91] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[92] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma,

et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision (IJCV)*, 123:32–73, 2017.

[93] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019.

[94] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10294–10303, 2019.

[95] Chao Lou, Wenjuan Han, Yuhuan Lin, and Zilong Zheng. Unsupervised vision-language parsing: Seamlessly bridging visual scene graphs with language structures via dependency relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15607–15616, 2022.

[96] Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Transformation driven visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6903–6912, 2021.

[97] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2017.

[98] Will Norcliffe-Brown, Efstathios Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. *arXiv preprint arXiv:1806.07243*, 2018.

[99] Weixin Liang, Yanhao Jiang, and Zixuan Liu. Graghvqa: Language-guided graph neural networks for graph-based visual question answering. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 79–86, 2021.

[100] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[101] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[102] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

[103] Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y Hammerla. Relational graph attention networks. *arXiv preprint arXiv:1904.05811*, 2019.

[104] Parth Shah, Prateek Shenoy, Shivansh Bhattad, Prathamesh Bhingardive, Abir Chakraborty, and Subbarao Kambhampati. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8876–8884, 2019.

[105] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, and Devi Parikh. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4602–4612, 2019.

[106] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019.

[107] Maryam Ziaeefard and Freddy Lécué. Towards knowledge-augmented visual question answering. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, page 1863–1873, 2020.

[108] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 489–498, 2020.

[109] Qingxing Cao, Bailin Li, Xiaodan Liang, Keze Wang, and Liang Lin. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2758–2767, 2022.

[110] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2712–2721, 2022.

[111] Mingxiao Li and Marie-Francine Moens. Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[112] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5089–5098, 2022.

[113] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning (ICML)*, 2023.

[114] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:37309–37323, 2022.

[115] Hongyu Ren, Hanjun Dai, Bo Dai, Xinyun Chen, Michihiro Yasunaga, Haitian Sun, Dale Schuurmans, Jure Leskovec, and Denny Zhou. Lego: Latent execution-guided reasoning for multi-hop question answering on knowledge graphs. In *International conference on machine learning*, pages 8959–8970, 2021.

[116] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3081–3089, 2022.

[117] Yanan Wang, Jianming Wu, Kazuaki Furumai, Shinya Wada, and Satoshi Kurihara. Vae-based adversarial multimodal domain transfer for video-level sentiment analysis. *IEEE Access*, 10:51315–51324, 2022.

[118] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022.

[119] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, 2021.

[120] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[121] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[122] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning (ICML)*, pages 448–456, 2015.

[123] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[124] Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Coarse-to-fine reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4558–4566, 2022.

[125] Shreya Ghosh, Abhinav Dhall, Nicu Sebe, and Tom Gedeon. Predicting group cohesiveness in images. In *International Joint Conference on Neural Networks (IJCNN)*, 2019.

[126] H. Hung and D. Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 12(6), 2010.

[127] Garima Sharma, Shreya Ghosh, and Abhinav Dhall. Automatic group level affect and cohesion prediction in videos. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 161–167, 2019.

[128] Terrence Fong, Charles Thorpe, and Charles Baur. Collaboration, dialogue, human-robot interaction. In *Robotics Research*, 2003.

[129] Tien Xuan Dang, Soo-Hyung Kim, Hyung-Jeong Yang, Guee-Sang Lee, and Thanh-Hung Vo. Group-level cohesion prediction using deep learning models with a multi-stream hybrid network. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 572–576, 2019.

[130] Da Guo, Kai Wang, Jianfei Yang, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. Exploring regularizations with face, body and image cues for group cohesion prediction. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, 2019.

[131] Bin Zhu, Xin Guo, Kenneth Barner, and Charles Boncelet. Automatic group cohesiveness detection with multi-modal features. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, 2019.

[132] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[133] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[134] Xin Huang, Yuxin Peng, and Mingkuan Yuan. Cross-modal common representation learning by hybrid transfer network. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1893—1900, 2017.

[135] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of vision*, 2007.

[136] Abhinav Dhall, Jyoti Joshi, Karan Sikka, Roland Goecke, and Nicu Sebe. The more the merrier: Analysing the affect of a group of people in images. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2015.

[137] Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9841–9850, 2019.

[138] Kristen A Lindquist, Jennifer K MacCormack, and Holly Shablack. The role of language in emotion: Predictions from psychological constructionism. *Frontiers in Psychology*, 6:444, 2015.

[139] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[140] An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[141] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[142] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.

[143] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

[144] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5561–5570, 2018.

[145] Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. Entity-aware image caption generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4013–4023, 2018.

[146] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

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

[148] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[149] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8807–8817, 2019.

[150] P. Tzirakis, J. Zhang, and B. W. Schuller. End-to-end speech emotion recognition using deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093, 2018.

[151] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *The annual conference of the International Speech Communication Association (Interspeech)*, pages 2803–2807, 2019.

[152] Xin Guo, Bin Zhu, Luisa F Polanía, Charles Boncelet, and Kenneth E Barner. Group-level emotion recognition using hybrid deep models based on faces,

scenes, skeletons and visual attentions. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 635–639, 2018.

[153] Kai Wang, Xiaoxing Zeng, Jianfei Yang, Debin Meng, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. Cascade attention networks for group emotion recognition with face, body and image cues. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 640–645, 2018.

[154] Lorenzo Magnani, Sabino Civita, and Guido Previde Massara. Visual cognition and cognitive modeling. In *Human and Machine Vision*, pages 229–243, 1994.

[155] Jessica A Collins and Ingrid R Olson. Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic bulletin & review*, 21(4):843–860, 2014.

[156] Steven Pinker. Visual cognition: An introduction. *Cognition*, 18(1-3):1–63, 1984.

[157] Richard GM Morris, Lionel Tarassenko, and Michael Kenward. *Cognitive Systems-Information Processing Meets Brain Science*. Elsevier, 2005.

[158] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 1459–1462, 2010.

[159] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[160] Susmitha Vekkot, Deepa Gupta, Mohammed Zakariah, and Yousef Ajami Alotaibi. Hybrid framework for speaker-independent emotion conversion using i-vector plda and neural network. *IEEE Access*, 7:81883–81902, 2019.

[161] Sunan Li, Wenming Zheng, Yuan Zong, Cheng Lu, Chuangao Tang, Xingxun Jiang, Jiateng Liu, and Wanchuang Xia. Bi-modality fusion for emotion recognition in the wild. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 589–594, 2019.

[162] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 292–301, 2018.

[163] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[164] Björn W Schuller, Anton Batliner, Christian Bergler, Florian B Pokorny, Jarek Krajewski, Margaret Cychosz, Ralf Vollmann, Sonja-Dana Roelen, Sebastian Schnieder, Elika Bergelson, et al. The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. In *The annual conference of the International Speech Communication Association (Interspeech)*, pages 2378–2382, 2019.

[165] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1145, 1995.

[166] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021.

[167] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. *arXiv preprint arXiv:2005.03545*, 2020.

[168] Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and Alberto Del Bimbo. Search-oriented micro-video captioning. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 3234–3243, 2022.

[169] Liangli Zhen, Peng Hu, Xi Peng, Rick Siow Mong Goh, and Joey Tianyi Zhou. Deep multimodal transfer learning for cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):798–810, 2022.

[170] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9180–9192, 2021.

[171] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7837–7851, 2022.

[172] Donghuo Zeng, Yi Yu, and Keizo Oyama. Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(3), 2020.

[173] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10386–10395, 2019.

[174] Ning Han, Jingjing Chen, Guangyi Xiao, Yawen Zeng, Chuhao Shi, and Hao Chen. Visual spatio-temporal relation-enhanced network for cross-modal text-video retrieval. *arXiv preprint arXiv:2110.15609*, 2021.

[175] Donghuo Zeng, Yanan Wang, Jianming Wu, and Kazushi Ikeda. Complete cross-triplet loss in label space for audio-visual cross-modal retrieval. In *IEEE International Symposium on Multimedia (ISM)*, pages 1–9, 2022.

[176] Li He, Xing Xu, Huimin Lu, Yang Yang, Fumin Shen, and Heng Tao Shen. Unsupervised cross-modal retrieval through adversarial learning. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1153–1158, 2017.

[177] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 154–162, 2017.

[178] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2018.

[179] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2827–2836, 2016.

[180] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *IEEE International Conference on Image Processing (ICIP)*, pages 6–10, 2019.

[181] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 4163–4174, 2020.

[182] Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. Dynabert: Dynamic bert with adaptive width and depth. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9782–9793, 2020.

[183] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10870–10879, 2020.

[184] Zhishan Li, Ying Nie, Kai Han, Jianyuan Guo, Lei Xie, and Yunhe Wang. A transformer-based object detector with coarse-fine crossing representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:38733–38746, 2022.

[185] Licheng Jiao, Jie Gao, Xu Liu, Fang Liu, Shuyuan Yang, and Biao Hou. Multiscale representation learning for image classification: A survey. *IEEE Transactions on Artificial Intelligence*, 4(1):23–43, 2023.

[186] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–18, 2022.

[187] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[188] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020.

[189] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3048–3068, 2022.

[190] Yanan Wang, Jianming Wu, Panikos Heracleous, Shinya Wada, Rui Kimura, and Satoshi Kurihara. Implicit knowledge injectable cross attention audiovisual model for group emotion recognition. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 827–834, 2020.

[191] Yanan Wang, Jianming Wu, Jinfa Huang, Gen Hattori, Yasuhiro Takishima, Shinya Wada, Rui Kimura, Jie Chen, and Satoshi Kurihara. Ldnn: Linguistic knowledge injectable deep neural network for group cohesiveness understanding. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 343–350, 2020.

[192] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on CLIP via vision-language knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL*, pages 2383–2395, 2022.

[193] Nikhil Rasiwasia, Dhruv Mahajan, Vijay Mahadevan, and Gaurav Aggarwal. Cluster canonical correlation analysis. In *AISTATS*, pages 823–831, 2014.

[194] Donghuo Zeng, Yi Yu, and Keizo Oyama. Audio-visual embedding for cross-modal music video retrieval through supervised deep cca. In *IEEE International Symposium on Multimedia (ISM)*, pages 143–150, 2018.

[195] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2), 2020.

[196] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3550–3558, 2018.

[197] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(1), 2020.

[198] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–536, 2019.

[199] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 2008.

[200] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2359–2369, 2020.

[201] Yi Yu, Suhua Tang, Kiyoharu Aizawa, and Akiko Aizawa. Category-based deep cca for fine-grained venue discovery from multimodal data. *IEEE Transactions on Neural Networks and Learning Systems*, 30(4):1250–1258, 2019.

[202] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[203] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284, 2018.

[204] Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44, 2018.

[205] Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2773–2785, 2019.

[206] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*, 2022.