論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Empowering Emotion Recognition with Flexible Modality Information
著者(和文)	東遠 李
Author(English)	Dongyuan Li
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12860号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:奥村 学,中山 実,鈴木 賢治,篠﨑 隆宏,船越 孝太郎
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12860号, Conferred date:2024/9/20, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

EMPOWERING EMOTION RECOGNITION WITH FLEXIBLE MODALITY INFORMATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF INFORMATION AND COMMUNICATION ENGINEERING
OF Tokyo Institute of Technology
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF ENGINEERING

Dongyuan Li
August 2024

Abstract

Teaching machines to understand human emotion is one of the most elusive and long-standing challenges in Artificial Intelligence. This thesis tackles two core issues of emotion recognition: (1) how to effectively and efficiently apply unimodal emotion recognition tools in real-world scenarios; (2) how to build a general multimodal emotion recognition model with high performance. Specifically, we focus on unimodal and multimodal emotion recognition: a class of emotion recognition models built on top of deep neural networks. Compared to traditional sparse, hand-designed feature-based machine learning methods or statistic models, these end-to-end neural models have proven to be more effective in learning and extracting rich sentiment and semantic information and improved performance on all modern emotion recognition benchmarks by a large margin.

This thesis consists of two parts. In the first part, we aim to cover the essence of unimodal emotion recognition and present our efforts at building effective and efficient unimodal emotion recognition models. Specifically, existing unimodal emotion recognition methods often overlook the information gap between the pre-trained models and the downstream emotion recognition task, resulting in sub-optimal performance. Moreover, current methods require much time for fine-tuning on each specific unimodal dataset, which limits their effectiveness in real-world scenarios with large-scale noisy data. To address these issues, we take speech emotion recognition as an example, and propose an active learning (AL)-based fine-tuning framework for speech emotion recognition, called AFTER, that leverages task adaptation pre-training (TAPT) and AL methods to enhance performance and efficiency. Specifically, we first use TAPT to minimize the information gap between the pre-trained speech recognition task and the downstream speech emotion recognition task. Then, AL methods are employed to iteratively select a subset of the most informative and diverse

samples for fine-tuning, thereby reducing time consumption. Experiments demonstrate that our method AFTER, using only 20% samples, improves precision by 8.45% and reduces time consumption by 79%. The additional extension of AFTER and ablation studies further confirms its effectiveness and applicability to various real-world scenarios. We also summarize limitations and discuss future directions in this field.

In the second part of this thesis, we aim to cover the essence of multimodal emotion recognition and present our efforts at building effective and robust multimodal emotion recognition models. Specifically, graph-based multimodal emotion recognition models have achieved state-of-the-art performance on multiple benchmarks. However, current graph-based methods fail to simultaneously depict global contextual features and local diverse unimodal features in a dialogue. Furthermore, with the number of graph layers increasing, they easily fall into over-smoothing. In this paper, we propose a method for **jo**int modalit**y** **fu**sion and graph contrastive **l**earning for multimodal emotion recognition (JOYFUL), where multimodal fusion, contrastive learning, and emotion recognition are jointly optimized. Specifically, we first design a new multimodal fusion mechanism that can provide deep interaction and fusion between the global contextual and unimodal specific features. Then, we introduce a graph contrastive learning framework with inter-view and intra-view contrastive losses to learn more distinguishable representations for samples with different sentiments. Extensive experiments on three benchmark datasets indicate that JOYFUL achieves state-of-the-art performance compared to all baselines. We also summarize recent advances and discuss future directions and open questions in this field.

Acknowledgments

The past three years at Tokyo Institute of Technology have been an unforgettable and invaluable experience for me. When I first started my Ph.D. in 2021, I could not speak Japanese and English fluently and was a true beginner in the field “natural language processing”. It is unbelievable that over the following years I have actually been doing research about language and training computer systems to understand human languages, as well as training myself to speak and write in English and Japanese. At the same time, 2022 is the year that large language models started to take off and dominate almost all the AI applications we are seeing today. I witnessed how fast Generative Artificial Intelligence has been developing from the beginning of the journey and feel quite excited and occasionally panicked to be a part of this trend. I would not have been able to make this journey without the help and support of many, many people and I feel deeply indebted to them.

First of all, my greatest thanks go to my advisor, Okumura Manabu. I really didn’t know Okumura sensei when I first came to Tokyo Institute of Technology— only after a couple of years that I worked with him and learned about NLP, did I realize how privileged I am to get to work with one of the most brilliant minds in our field. During the process of expressing gratitude to him, many scenes came to mind. During 2022 to 2023, my paper was repeatedly rejected by conferences such as IJCAI, ACL, and ICASSP, and encountered significant setbacks. Okumura sensei gave me great encouragement and repeatedly helped me improve the quality of my academic paper and rebuild my academic confidence. In the end, all the rejected papers were successfully accepted. He is always able to quickly and accurately identify my academic loopholes and shortcomings. I really enjoy discussing and exchanging ideas with Okumura sensei on how to revise papers, which greatly enhances my academic taste and ability. I often trouble him to help me revise my paper late at night,

and he never complains. He always unconditionally supports us in attending academic conferences abroad and exchange ideas with world-class scholars. He introduced me to many excellent professors and gave me the opportunity to visit the National University of Singapore and the University of Cambridge. Okumura sensei's academic attitude is worth learning for me for a lifetime. I hope to become a teacher like Okumura sensei in the future and pass his help from me to more students.

I would like to thank Funakoshi Kataro sensei and Kosugi Satoshi sensei— the other two giants of the Okumura-Funakoshi NLP group — for being on my thesis committee and for a lot of guidance and help throughout my Ph.D. studies. Funakoshi sensei is also my academic supervisor during my doctoral period. Funakoshi sensei gave me a lot of useful suggestions on the seminar, which helped me improve my research motivation. He often helps me revise my paper word for word, sentence for sentence, design experiments, and provide valuable feedback from the perspective of a reviewer. His rigorous attitude towards academia and meticulous spirit remind me of the craftsmanship spirit of Japan. Funakoshi sensei taught me the most precious quality during my studies in Japan, which is to meticulously conduct academic research, ensure that all work is done to the fullest, and have a clear conscience. And for Kosugi sensei, although I haven't known Kosugi Sensei for a long time, his help to me has been tremendous. We sat in the same room, and I could always ask him questions at any time. He would always put down his busy work to discuss with me, and our discussions always gave me a lot of inspiration. He will do his best to help me revise my paper, and I can always gain huge benefits from his revisions.

It is also my great honor to have Prof. Takahiro Shinozaki, Prof. Kenji Suzuki, Prof. Kataro Funakoshi and Prof. Minoru Nakayama on my thesis committee. I am very grateful to them for taking the time and effort to participate in my doctoral defense. They have given me many valuable opinions, whether it is the issues in presentation or the shortcomings in research, which have benefited me greatly.

During 2024, I have done one wonderful research internship at the University of Tokyo. I thank my mentor Renhe Jiang, who has provided tremendous help to my research. We got to know each other through participating in the CIKM conference held in the UK. We discovered that we come from the same city and university in China, and through gradual communication, our friendship has deepened. With his help, we have conducted a lot of

research and I am very grateful for his help. I am also very grateful to Professor Min-Yen Kan from the National University of Singapore and Professor Simone Teufel from the University of Cambridge for leading me on a campus tour and providing valuable feedback on my research topic. I am also very grateful to Professor Hidetaka Kamigaito, Professor Hiroya Takamura, Professor Xiaoke Ma, Professor Ding Nan, and Professor Weiping Ding for providing valuable feedback on my research content.

I thank the whole Okumura-Manabu NLP Group, especially Iiyama san, Aru Maekawa, Thodsaporn Chayintr, Chenlong Hu, Dangwang Chen, Jian Wu, Jialun Shen, Kexin Ren, Yusong Wang, Xinran Shao, Shiyin Tan, Toshiki Kawamoto, Tianjiao Zhu, Yujun Chen, Yicheng Xu, Ying Zhang, Zhen Wang, Zifan Wang and others. Expecially, I really enjoy drinking coffee with Aru Maekawa every day to exchange academic ideas. He is a particularly intelligent NLP researcher who has given me a lot of academic help and advice. I am particularly grateful to Iiyama san for helping me reimburse my travel expenses every time. With her support, I have more time to focus on my academic pursuits. Thanks to Boat san, who takes me to the gym every day and is my fitness instructor. Thank you very much to Shiyin Tan, Zhen Wang, Ying Zhang and Yusong Wang for their help. We are both close collaborators and close friends. Yusong Wang provided me with tremendous help during my difficult times in life, I am very grateful for his help and support. I hope I can also be his support in difficult times. Thank you very much, Xinran Shao, Wangchen Dang, Yujun Chen, Kexin Ren, they comforted and encouraged me to accompany me during my emotional distress, and I am very grateful for their help. I also have many thanks to every member of the laboratory, hoping that everyone can achieve satisfactory results in their research and life.

Outside of the NLP group, I have been extremely lucky to be surrounded by many great friends. Just to name a few (and forgive me for not being able to list all of them): Xiangbo Li, Linze Li, and Hui Li, my close friend for many years, who keeps pulling me out of my stressful Ph.D. life, and I share a lot of joyous moments with them. Benhui Zhang, Zhihao Huang, Wenming Wu, Yu Feng, Chenxi Tian, Wei Zhao, Haiyue Wang, my classmate at Xidian University. Rui Zhang, Fengkai Li, Xiangyu Yao, my classmate at Dalian University of Technology, Runze Chen, and Yingxin Jiang, my friends in high school.

I thank my parents: Yu-E Duan and Yong-Ming Li. Like most Chinese students in my

generation, I am the only child of my family and I have a very close relationship with them — even if I can only spare 2–3 weeks staying with them every year. My parents made me who I am today and I never know how to pay them back. I hope they are at least a little proud of me for what I have been through so far.

To my parents for their unconditional love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

"Integration of information from multiple sensory channels is crucial to understand the tendencies and reactions of humans" (Partan and Marler, 1999). Emotion recognition aims exactly at identifying and tracking the emotional state of each utterance from heterogeneous visual, audio, and text channels (Li et al., 2023b; Lu et al., 2024). Due to its potential applications in human-computer interaction systems (Li et al., 2022c), social media analysis (Gupta et al., 2022), bioinformatics (Nicolson et al., 2023), and recommendation systems (Singh et al., 2022), emotion recognition has received increasing attention (Poria et al., 2021).

However, teaching machines to understand human emotion is one of the most elusive and long-standing challenges in Artificial Intelligence (Cowie et al., 2001). Before we proceed, *we must ask what it means to understand human emotion?* Figure 1.1 demonstrates a general pipeline of emotion recognition. To process such text, image, and audio signals, the NLP, CV, and Speech community has put decades of effort into solving different tasks for various aspects of emotion recognition, including:

(a) **Textual Emotion Recognition**. It requires our machines to understand the natural language. For example, in the sentence "My first publication made my palms sweat." machines should understand the meaning of "palms sweat" and predict the emotion as "nervous", "fear" and "embarrassment" instead of "happy" for this sentence.

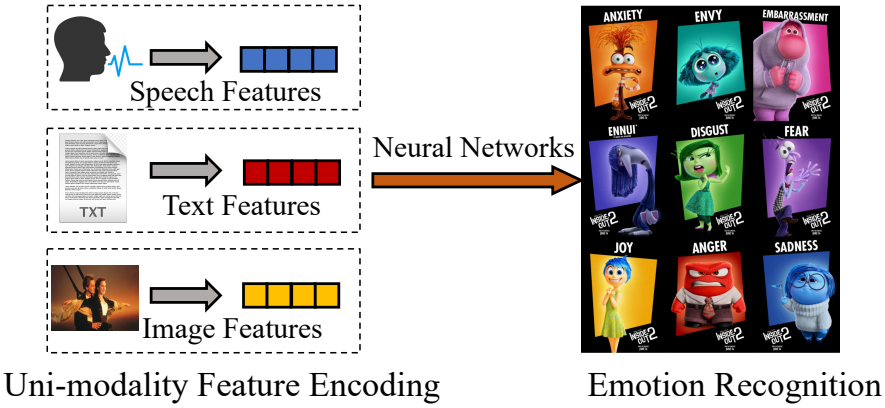


Figure 1.1: General pipeline of emotion recognition. The right side figure comes from one movie "Inside Out" from Pixar Animation Studios [2] that describes human emotions.

(b) **Facial Emotion Recognition**. Machines also need to understand human micro expressions and use them to understand and recognize human emotion. For example, as shown in Figure 1.2, machines need to first extract micro expressions from the images, which have discriminative features and can improve the accuracy of emotion recognition.

Figure 1.2: Facial emotion recognition. The figure comes from the website [3].

(c) **Speech Emotion Recognition**. Instead of only understanding the meaning of language, machines need to extract tones and intonations in speech for emotion recognition. For example, consider the speech "I would like to borrow this book from you". If it is a slow and deep speaking pace, it may display tense emotions, and if it is an excited and high pitched tone, it may display happy emotions.

(d) **Multimodal Emotion Recognition**. Due to the complementary between various single-mode information, machines need to reasonably fuse multiple unimodal specific features and effectively remove redundant features.

Although many methods have achieved considerable success, there are still many issues waiting to be addressed. The first and foremost issue is ***"How to effectively apply emotion recognition tools in real-world scenarios"***. To explore this issue, we consider the simplest unimodal emotion recognition, *i.e.*, speech emotion recognition.

Current methods for speech emotion recognition can be broadly classified into two categories: traditional machine learning-based methods and deep learning-based methods. Specifically, traditional machine learning-based methods typically consist of three main components: speech feature extraction, feature selection, and emotion recognition. However, selecting and designing features tailored to specific corpora often lack generalizability across other datasets and consume significant time. Deep learning-based methods can address these issues by automatically extracting more abstract features to improve generalization. With the development of pre-trained language models (Devlin et al., 2019) and the availability of large-scale datasets, various pre-trained automatic speech recognition models have been proposed. These automatic speech recognition models use speech's acoustic and linguistic properties to provide more robust and context-aware representations for speech signals. Xia et al. (2021) proved that fine-tuning wav2vec 2.0 (Schneider et al., 2019) on speech emotion recognition datasets achieves state-of-the-art performance on IEMOCAP (Baevski et al., 2020). This finding has inspired researchers to explore new fine-tuning strategies in automatic speech recognition models, becoming a new paradigm for speech emotion recognition. For example, Ren et al. (2022) proposed a self-distillation speech emotion recognition model to fine-tune wav2vec 2.0, obtaining state-of-the-art performance on the DEMoS dataset. Ferreira (2022) fine-tuned wav2vec 2.0 by jointly optimizing speech emotion recognition and automatic speech recognition tasks, achieving state-of-the-art performance in Portuguese datasets. However, several issues still need to be addressed.

(1) Current methods ignore the information gap between pre-trained models and downstream tasks. For example, pre-training automatic speech recognition model wav2vec 2.0 adopts the masked learning objective to predict missing frames from the remaining

context, while the downstream speech emotion recognition task aims to minimize cross-entropy loss between predicted and referenced emotion labels. Gururangan et al. (2020) proved that the information gap would decrease the performance of downstream tasks.

(2) Current methods lack generalization. For example, the best-performing STRFs (Xia et al., 2021) train their models on the IEMOCAP dataset, leading to poor generalization for unseen datasets. Real-world scenarios contain much heterogeneous and noisy data, which hinders the application of these speech emotion recognition methods. Specifically, outliers encompass various ambiguous emotions due to the complexity of speech, which can lead to inaccurate emotional annotations and degrade the performance of the model. Training redundant samples repeatedly does not improve the model accuracy. Instead, they lead to an uneven distribution of data, making it more challenging to identify emotion with a limited amount of data.

(3) Pre-trained automatic speech recognition models often contain millions of parameters, for example, wav2vec 2.0 contains 317 million parameters. Fine-tuning them for real-world tasks with large-scale datasets is time-consuming and not realistic.

Unimodal emotional datasets are always easy to obtain in the real world, while each unimodality can only describe one aspect of human complex emotions. Considering that audio, image, and text can provide complementary information for each other, multimodal emotion recognition methods have been proposed for better performance. However, ***"how to build a general multimodal emotion recognition model with high performance"*** is still an open and challenging issue for this study.

Specifically, Figure 1.3 shows that emotions expressed in a dialogue are affected by three main factors: (1) multiple unimodalities, *e.g.*, different modalities such as text, speech and image can complete each other to provide a more informative utterance representation; (2) global contextual information, *e.g.*, u_3^A depends on the topic "The ship sank into the sea", predicting fear as its emotion state; and (3) intra-person and inter-person dependencies, *e.g.*, u_6^A becomes sad affected by sadness in u_4^B and u_5^B. Depending on how to model intra-person and inter-person dependencies, current methods can be categorized into Sequence-based and Graph-based methods. Sequence-based models use recurrent neural networks or Transformers to model the temporal interaction between utterances (Dai et al., 2021;

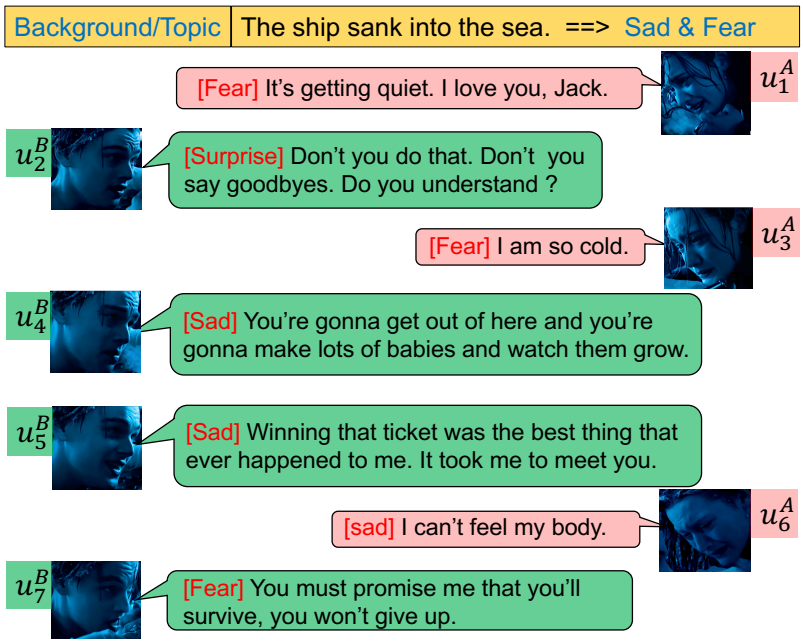


Figure 1.3: Emotions are affected by multiple unimodalities, global contextual, intra- and inter-person dependencies. Images are from the movie "Titanic".

Mao et al., 2022; Liang et al., 2022). However, they failed to distinguish intra-speaker and inter-speaker dependencies. Furthermore, they tend to easily lose unimodal specific features by the cross-modal attention mechanism (Rajan et al., 2022). The graph structure solves these problems by using the edges between nodes (speakers) to distinguish intra-speaker and inter-speaker dependencies (Joshi et al., 2022; Wei et al., 2019). Graph Neural Networks (GNNs) further help nodes learn common features by aggregating information from neighbors while maintaining their unimodal specific features.

Although graph-based multimodal emotion recognition methods have achieved great success, there still remain several problems that need to be solved:

(1) Current methods always directly aggregate features of multiple modalities (Joshi et al., 2022) or project modalities into a latent space to learn utterance representations (Li et al., 2022f), which ignores the diversity of each modality and fails to capture richer semantic information from each modality.

(2) Current studies ignore the importance of dynamically changing global contextual information during the feature fusion process, leading to poor performance.

(3) Almost all graph-based methods adopt GNN (Scarselli et al., 2009) or Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) as backbones for multimodal emotion recognition. However, with the number of layers deepening, the phenomenon of over-smoothing starts to appear. This phenomenon results in the representation of similar sentiments being indistinguishable.

(4) Most methods use a two-phase pipeline (Fu et al., 2021; Joshi et al., 2022), where they first extract and fuse unimodal features as utterance representations and then fix them as input for graph models. However, the two-phase pipeline will lead to sub-optimal performance since the fused representations are fixed and cannot be further improved to benefit from the downstream supervisory signals.

In summary, the motivation of this study is to address the following two issues: (1) design an efficient and effective method for unimodal emotion recognition and (2) design a high-performance general multimodal emotion recognition model.

1.2 Thesis Outline

This thesis consists of two parts. PART I UNIMODAL EMOTION RECOGNITION and PART II MULTIMODAL EMOTION RECOGNITION.

Specifically, **PART I** focuses on the task of unimodal emotion recognition, with a detailed introduction to current textual, facial, and speech emotion recognition. Furthermore, we provide an empirical study on speech emotion recognition.

In Chapter 2, we first give an overview of the history and recent development of the field of unimodal emotion recognition. Next, we formally define the problem formulation. We then briefly introduce the widely used unimodal datasets. Finally, we discuss the differences of two related tasks: emotion recognition and sentiment analysis.

In Chapter 3, to solve the issues that exist in current unimodal emotion recognition systems, we provide a new active learning-based task adaptation pre-training framework. Take

speech emotion recognition as a case study, we first introduce the previous best-performing speech emotion recognition methods and highlight the research motivation. Then, we introduce the most related work, *i.e.*, active learning and task adaptation pre-training to our framework. Next, we introduce our framework and how it can solve the issues in detail. Furthermore, we show the experimental results on multiple speech emotion recognition datasets, demonstrating the effective and efficiency of our proposed method. Finally, we introduce the limitations and future work of the proposed method and give the summary of this study.

PART II focuses on multimodal emotion recognition, with a detailed introduction to current multimodal fusion methods and our framework to solve the issues mentioned above in current multimodal emotion recognition work. Finally, we give the challenges and future work about multimodal emotion recognition.

In Chapter 4, we introduce the main multimodal fusion mechanisms of current studies. Then, we introduce the general backbones, datasets, and evaluation metrics of multimodal emotion recognition methods.

In Chapter 5, we propose a general graph contrastive learning-based multimodal emotion recognition framework. We first introduce the previous multimodal emotion recognition methods and highlight the research motivations of multimodal emotion recognition. Then, we propose joint multimodal fusion and graph contrastive learning for multimodal emotion recognition, where multimodal fusion, graph contrastive learning, and multimodal emotion recognition are jointly optimized in an overall objective function. Extensive experiments conducted on three multimodal benchmark datasets demonstrated the effectiveness and robustness of JOYFUL.

In Chapter 6, we introduce the challenges and future work from datasets and models. We raise several research questions waiting for future solutions.

In Chapter 7, we give the final conclusion of this thesis.

1.3 Contributions

For unimodal emotion recognition models, we have the following contributions.

(1) To the best of our knowledge, we are the first to propose a general task adaptation pre-training and active learning-based fine-tuning framework for the speech emotion recognition to address the information gap, noisy sensitive, and low efficiency issues.

(2) We created three additional large-scale speech emotion recognition datasets to simulate different complex real-world scenarios by merging existing high-quality speech emotion datasets. These datasets represent noisy and heterogeneous real-world situations. We released our datasets on Github [4] to share with other researchers.

(3) Extensive experiments demonstrate the effectiveness and efficiency of our method AFTER. It performs well on IEMOCAP, Merged Dataset with four emotional categories, as well as in the SAVEE and Merged-3 Dataset with seven emotional categories. Additional extensions of AFTER demonstrate the effectiveness and applicability.

For multimodal emotion recognition models, we have the following contributions:

(1) We propose a novel joint leaning framework for multimodal emotion recognition, where multimodal fusion, GCL, and emotion recognition are jointly optimized. Our new multimodal fusion mechanism can obtain better representations by simultaneously depicting global contextual and local unimodal specific features.

(2) To the best of our knowledge, JOYFUL is the first method to utilize graph contrastive learning for multimodal emotion recognition, which significantly improves the model's ability to distinguish different sentiments. Multiple graph augmentation strategies further improve the model's stability and generalization.

(3) We release our source code on Github [5]. Extensive experiments conducted on three multimodal benchmark datasets showed the effectiveness and robustness of our framework.

[4] https://github.com/Clearloveyuan/AFTER
[5] https://github.com/Clearloveyuan/MERC-main

Part I

Unimodal Emotion Recognition

Chapter 2

Overview of Unimodal Emotion Recognition

> Emotion recognition is the process of identifying human emotion. People vary widely in their accuracy at recognizing the emotions of others. Use of technology to help people with emotion recognition is a relatively nascent research area. Generally, the technology works best if it uses multiple modalities in context. To date, the most work has been conducted on automating the recognition of facial expressions from ***video***, spoken expressions from ***audio***, written expressions from ***text***, and physiology as measured by ***wearables***.
>
> Wikipedia: https://en.wikipedia.org/wiki/Emotion_recognition

In this chapter, we aim to provide readers with an overview of unimodal emotion recognition. Specifically, we begin with the history of emotion recognition in Section 2.1, from the early systems developed in the 1970s (Section 2.1.1), to attempts to build machine learning models for this task (Section 2.1.2), and to the more recent resurgence of neural approaches in Section 2.1.3, *i.e.*, textual emotion recognition in Section 2.1.3.1, speech emotion recognition in Section 2.1.3.2 and facial emotion recognition 2.1.3.3.

Then, we introduce the emotion recognition assessment metrics in Section 2.2.

Next, we introduce the main datasets in Section 2.3. Specifically, we introduce the main

datasets for textual emotion recognition in Section 2.3.1, speech emotion recognition in Section 2.3.2 and facial emotion recognition in Section 2.3.3.

Finally, we discuss the difference between emotion recognition and sentiment analysis in Section 2.4.

2.1 Previous Studies

2.1.1 Early Knowledge-based Systems

The history of building automated emotion recognition systems dates back more than 40 years ago. In the 1970s, researchers already recognized the importance of emotion recognition as an appropriate way of understanding the behavioral patterns of human individuals and groups (Shimoda et al., 1978) and human diseases, such as schizophrenics (Walker et al., 1980) and emotionally disturbed children (Zabel, 1979).

One of the most notable early works is the SO-CAL system detailed in Taboada et al. (2011). SO-CAL is a lexicon-based approach that uses dictionaries of words annotated with their semantic orientation (polarity and strength) and incorporates intensification and negation. It is applied to the polarity classification task, which involves assigning a positive or negative label to a text based on the opinion expressed toward its main subject matter. This work set a strong vision for emotion recognition, but the actual system was very small, limited to hand-coded scripts, and difficult to generalize to broader domains.

To address the issues mentioned above, many works use broader knowledge-based resources during the emotion classification process, such as WordNet [1], SenticNet (Cambria et al., 2022), ConceptNet [2], and EmotiNet (Balahur et al., 2012). Specifically, WordNet is a lexical database of semantic relations between words that links words into semantic relations, including synonyms, hyponyms, and meronyms. The synonyms are grouped into synsets with short definitions and usage examples. Many works check whether sentences contain vocabulary related to emotion based on WordNet to determine emotions (Badaro et al., 2018; Kocon, 2023) in sentences. One of the advantages of this approach is the

[1] https://wordnet.princeton.edu/

[2] https://en.wikipedia.org/wiki/Open_Mind_Common_Sense#ConceptNet

accessibility and economy brought about by the large availability of such knowledge-based resources. A limitation of this technique, on the other hand, is its inability to handle concept nuances and complex linguistic rules (Cambria, 2016).

In summary, knowledge-based techniques can be mainly classified into two categories: dictionary-based and corpus-based approaches (Darwich et al., 2019). Dictionary-based approaches find opinion or emotion seed words in a dictionary and search for their synonyms and antonyms to expand the initial list of opinions or emotions (Madhoushi et al., 2015). Corpus-based approaches, on the other hand, start with a seed list of opinion or emotion words and expand the database by finding other words with context-specific characteristics in a large corpus. Although corpus-based approaches take into account context, their performance still varies in different domains since a word in one domain can have a different orientation in another domain (Hemmatian and Sohrabi, 2019).

2.1.2 Machine Learning Approaches

Statistical methods commonly involve the use of different supervised machine learning algorithms in which a large set of annotated data is fed into the algorithms to learn and predict the appropriate emotion types. Machine learning algorithms generally provide more reasonable classification accuracy compared to other approaches, but one of the challenges in achieving good results in the classification process is the need to have sufficiently large training datasets. Different unimodal emotion recognition works always have different machine learning approaches. In this study, to simplify this part, we take speech emotion recognition (SER) as an example to introduce its previous work. In SER, feature engineering and designing ML models for classification or prediction are often considered separate problems. Most of the actual SER research has focused on feature engineering or the design of pre-processing data transformation pipelines to craft emotional representations that support ML algorithms. Although feature engineering techniques can help improve the SER performance, the downside is that these techniques are labor intensive and time-consuming. For decades, Mel frequency cepstral coefficients (MFCCs) (Furui, 1986) have been used as the main set of features for SER and other speech analysis tasks. The four steps involved in the extraction of MFCCs are: (1) computation of the Fourier transform, (2) projection

of the powers of the spectrum onto the Mel scale, (3) taking the logarithm of the Mel frequencies, and (4) applying discrete cosine transformation or other suited transformations for compressed representations. The last step is found to lose information and destroy spatial relations; therefore, it is usually omitted, resulting in the LogMel spectrum, a popular feature used by the speech community. It is also the most popular feature to train DL networks in the speech domain. Minimalist feature sets such as GeMAPs and eGeMAPs (Eyben et al., 2016) are also widely used as benchmarks. They are designed/engineered to (a) index affective physiological changes in voice production and (b) achieve automatic extractability.

2.1.3 Deep Learning Approaches

Deep learning, which is part of the unsupervised family of machine learning, is also widely used in emotion recognition. Well-known deep learning algorithms include different architectures of artificial neural networks (ANN), such as the convolutional neural network (CNN), long-short-term memory (LSTM) and the extreme learning machine (ELM). The popularity of deep learning approaches in the domain of emotion recognition may be mainly attributed to its success in related applications such as computer vision, speech recognition, and Natural Language Processing (NLP). In this section, we will introduce the mainstream methods of text, speech, and image emotion recognition methods in order.

2.1.3.1 Textual Emotion Recognition Models

Textual emotion recognition (TER) aims to classify a textual expression into one or several emotion categories, depending on the underlying emotion theories. Inspired by the successful transfer learning of CNNs from the image field to other computer vision fields, the emergence of a pre-trained language model opened the pre-training era in the NLP field. They generate contextualized word embedding with general knowledge that can be easily transferred to almost all downstream tasks. We summarize current TER methods according to their representation methods, including word-level emotional representation methods, multi-feature fused emotional representation methods, and knowledge-enhanced emotional representation methods. We will introduce them sequentially in order.

Word-Level Emotional Representation Methods. To learn generalized word-level emotion representation, Emo2Vec (Xu et al., 2018) is designed to encode emotional semantics in vector representations, facilitating the learning of generalized token-level emotion representations. This model is pre-trained using multi-task learning on six emotion-related tasks, making it versatile in various applications. Emo2Vec is often combined with other word embeddings, which improves performance in emotional recognition. Emojis, widely used in digital communication to express emotions, play a significant role in popular culture. DeepMoji (Felbo et al., 2017) generates rich emotional representations by mapping emojis to continuous vectors. It is pre-trained on 1.2 billion tweets using a two-layer BiLSTM network with an attention mechanism, capturing diverse emotional nuances. Another approach, the domain-sensitive and sentiment-aware embedding model proposed by Shi et al. (2018), integrates sentiment semantics and domain specificity of words. This model further enhances the ability to recognize emotions accurately. These models incorporate emotional information into word embeddings, significantly improving emotional recognition performance.

Multi-Feature Fused Emotional Representation. In Ying et al. (2019), feature fusion is achieved by combining general knowledge from a fine-tuned BERT model with domain-specific knowledge from a convolutional network. The experimental results emphasize the critical role of domain knowledge in domain-specific applications. The architecture proposed in Meisheri and Dey (2018) uses two parallel attention LSTM towers to focus on encoding emotion-specific words. Two types of word embeddings are fed into these towers separately. The outputs undergo feature fusion and max-pooling to extract the most prominent features. In Jain et al. (2017), an ensemble model is proposed to predict the intensity of four emotions: anger, fear, joy, and sadness. The model comprises three individual models: feed-forward neural networks, multi-task learning networks, and CNN-LSTM-based networks. The final prediction is a weighted average of the outputs of these individual models, achieving the best performance in the WASSA 2017 shared task on emotion intensity. The work in Perikos and Hatzilygeroudis (2016) constructs an ensemble model using three classifiers: a knowledge-based classifier and two statistical classifiers (Naive Bayes and Maximum Entropy learner). The final prediction is determined by a majority voting approach. In Corchs et al. (2019), five independent models are trained on image and textual features, including Naive Bayes,

Bayesian Network, Decision Tree, KNN, and SVM. These models are then combined using the Bayesian model averaging method (Fersini et al., 2014).

Knowledge Enhanced Emotional Representation. With the development of deep learning networks, integrating prior knowledge as auxiliary information has become essential for the Text Emotion Recognition task. Prior knowledge includes resources in the emotional lexicon, common sense, linguistic patterns, and any other information related to emotions. This integration enhances the emotional feature representation of textual data, leading to a deeper understanding of emotions. Deep learning-based features are often combined with lexicon-based features to effectively introduce emotional resources into deep learning networks. These combined features can be fed into deeper networks to generate high-level, abstract features or directly into classifiers for final prediction. For example, in Khanpour and Caragea (2018), lexicon-based features were combined with CNN outputs and then fed into an LSTM network to detect emotion states from health-related posts. Similarly, in Akhtar et al. (2018), features extracted from an intermediate layer of a pre-trained network were concatenated with hand-crafted feature vectors, such as TF-IDF weighted word vectors and lexicon-based features. To address issues such as misspellings and words out of the vocabulary, some works combine lexical features with neural features to enhance performance (Agrawal and Suri, 2019). Incorporating lexicon-based emotional knowledge helps capture hidden semantics, providing a more insightful understanding of emotional texts. Enhancing word-level representation with external knowledge facilitates inference of implicit emotions through rich commonsense information. In Kumar et al. (2018), a two-layer attention network enriched with knowledge is proposed. This network applies primary word-level attention to input words and related terms from WordNet and Distributed Thesaurus, generating word embeddings enhanced by the knowledge graph. Secondary attention at the sentence level further refines context learning, leading to remarkable performance on the SemEval 2017 Task 5 benchmark. Rule-based representations are also widely accepted. Rule-embedded neural networks (ReNN) proposed in Wang (2018) encode domain knowledge and commonsense information, reducing computational complexity and improving model training efficiency with smaller datasets.

2.1.3.2 Speech Emotion Recognition Models

Speech emotion recognition (*SER*) task is one of the key components of human-machine interaction and human communication systems (Latif et al., 2023). With the development of deep learning, several attempts have been made to automatically learn emotion representations from audio signals using neural networks (Chang et al., 2023; Chen et al., 2023b; Dang et al., 2023). Current studies can be classfied into three categories: supervised SER, semi-supervised SER, and self-supervised SER. We will introduce them in order.

Supervised Speech Emotion Recognition. In Speech Emotion Recognition (SER), supervised representation learning methods are extensively utilized to enhance performance. For example, Huang et al. (2014a) employed a Deep Belief Network (DBN) to learn emotional representation from speech, achieving a 7% higher classification accuracy (86.5%) in the BUAA emotional corpus compared to traditional hand-engineered features. To further improve SER performance, Zou et al. (2016) integrated classical features with emotional representations learned by a DBN. Their findings demonstrated that fusing deep emotional representations with classical features improved the SER accuracy by 8.8%. Similarly, Latif et al. (2018) conducted experiments on multiple datasets, showing that DBNs could learn more powerful and effective long-range discriminative features, significantly improving the performance of SER. In addition to DBNs, researchers have also explored Deep Neural Networks (DNNs) with multiple fully connected hidden layers to learn emotional representation. Han et al. (2014) utilized DNNs to learn high-level emotional representations from raw speech. They constructed utterance-level representations from segment-level probability distributions produced by the DNN and used extreme learning machines for emotion classification on these utterance-level representations. Their proposed framework, evaluated on the IEMOCAP dataset, demonstrated a 20% improvement in classification accuracy, effectively capturing emotional representations. Moreover, Mirsamadi et al. (2017) evaluated various BLSTM-RNN architectures for learning emotional representations from speech. Using the IEMOCAP corpus for evaluation, they found that RNNs could learn both short-term frame-level acoustic representations and compact utterance-level emotional representations. Their results indicated that BLSTMs outperformed DNNs and SVMs trained on hand-engineered

features, underscoring the effectiveness of RNNs in capturing nuanced emotional information in speech. These studies collectively highlight the significant advancements in SER achieved through supervised representation learning methods, demonstrating the superiority of deep learning approaches over traditional feature engineering techniques.

Semi-supervised Speech Emotion Recognition. Obtaining large labeled datasets or pre-trained networks for specific problems like SER is not always feasible. Annotating emotional speech data is particularly challenging, expensive, and time-consuming, as it requires manual expert efforts. Semi-supervised representation learning addresses this issue by leveraging feature representations from large amounts of unlabeled data in combination with labeled data to build more effective classifiers. This approach significantly reduces the need for extensive human annotation and typically achieves higher accuracy compared to unsupervised representation learning. For example, Huang et al. (2014b) employ a CNN in a semi-supervised manner to capture emotional representations. Their model, evaluated on four publicly available datasets, demonstrates that a semi-supervised CNN can learn salient, orthogonal, and discriminative representations for SER. These representations lead to superior performance compared to traditional hand-engineered features. Deng et al. (2018) propose a semi-supervised model that combines an Autoencoder (AE) with a classifier. They treat samples from unlabeled data as an additional "garbage" class in the classification task. Evaluations on five publicly available datasets show that the features learned by this semi-supervised AE improve SER performance over those learned by an unsupervised AE. Parthasarathy and Busso (2018) employ a ladder network with skip connections between the encoder and decoder networks for SER. Their evaluation on the MSP-Podcast dataset demonstrates that a semi-supervised ladder network can learn more powerful representations, leading to better performance in predicting emotional attributes compared to conventional Denoising Autoencoders (DAEs). Latif et al. (2022) leverage a semi-supervised ladder network to generate robust feature representations. This is achieved by simultaneously minimizing the sum of supervised classification and unsupervised cost functions. The features produced by the ladder network are then used as emotional representations for classification with an SVM. Evaluations of the IEMOCAP corpus reveal that the proposed framework improves performance by 2.6% compared to a DAE and by 5. 3% on static acoustic features.

Tao et al. (2019) employ a ladder network to generate emotional representations, performing experiments on the IEMOCAP dataset. Their findings indicate that this approach improves classification performance even with a limited number of labeled samples, outperforming DAEs, VAEs, and hand-engineered features. Sahu et al. (2022) utilize a conditional GAN to model feature representations and generate new data samples. Evaluations of the IEMOCAP and MSP-IMPROV datasets demonstrate that synthetic feature vectors can enhance SER performance in various settings. Furthermore, Zhao et al. (2020) introduce robust semi-supervised GANs to address the challenge of limited labeled data. Their model, evaluated on four publicly available datasets, effectively captures underlying emotional representations from both labeled and unlabeled data. The results show that their approach surpasses state-of-the-art supervised and semi-supervised models. Sahu et al. (2017) evaluate semi-supervised Adversarial Autoencoders (AAEs) for encoding emotional representations in compressed form and generating synthetic data samples. Their experiments on the IEMOCAP dataset reveal that AAEs can encode emotional representations without losing class discriminability and generate synthetic samples that augment the training data, thus improving SER performance. These studies highlight the effectiveness of semi-supervised learning techniques, particularly ladder networks and GANs, in enhancing emotional representation learning for SER. Using both labeled and unlabeled data, these approaches provide robust solutions to improve the accuracy and reliability of emotion classification in speech.

Self-supervised Speech Emotion Recognition. Commonly used SER datasets, such as MSP-Podcast (Srinivasan et al., 2022), IEMOCAP (Busso et al., 2008) and CMU-MOSEI (Zadeh et al., 2018b), are relatively small compared to automatic speech recognition datasets. This limitation restricts the ability for pre-trained automatic speech recognition models to improve the accuracy of emotion recognition. Self-supervised pre-trained models, such as Transformers, provide a solution by first learning from a large-scale speech corpus without explicit labeling (Baruah and Banerjee, 2022; Dissanayake et al., 2022). The knowledge learned from the pre-training datasets can then be transferred to downstream tasks by either using the model as a feature extractor (Lavania et al., 2023) or directly fine-tuning the entire model (Chen and Yu, 2023). Although initially introduced for natural

language processing (NLP), several SSL-based [3] pre-trained automatic speech recognition models have been developed for speech processing, including wav2vec 2.0 (Baevski et al., 2020), HuBERT (Mohamed and Aly, 2021), and CLAP (Wu et al., 2023). Taking wav2vec 2.0 (Baevski et al., 2020) as an example, which serves as the base model in this draft, it comprises a multi-layer convolutional neural network (CNN) designed to predict future frames based on past frames, achieving through the minimization of a contrastive loss. Additionally, wav2vec 2.0 utilizes a transformer-based architecture, employing a masked learning objective to predict missing frames within the given context. These pre-trained models consistently demonstrate state-of-the-art performance across various SER datasets. For example, Boigne et al. (2020) observe that the wav2vec 2.0 features surpass the traditional spectral-based features in SER applications. L.Chen and A.Rudnicky (2022) showcase the advantages of task-adaptive pre-training in the wav2Vec 2.0 model, leading to a significant improvement in overall model performance. Furthermore, Xia et al. (2021) performed a comparative analysis of features extracted with different temporal spans, concluding that features with longer temporal context, such as wav2vec 2.0, exhibit superior performance in SER. Pepino et al. (2021) demonstrate that the features derived from a linear combination of layers outperform single-layer representations in wav2vec 2.0 for SER applications.

2.1.3.3 Facial Emotion Recognition Models

Facial emotional recognition (FER) is the procedure of recognizing human emotion through verbal expressions, facial expressions, body movements, multiple physiological signals, and facial expressions. Usually, a basic FER program consists of two main steps: facial abstraction and facial expression recognition. Current studies can be mainly classfied into static image-based FER networks and dynamic image-based FER networks. We will introduce them in order.

Static Image-Based FER Networks. Many recent studies focus on expression recognition tasks using static images, largely due to the ease of statistical analysis and the availability of relevant training and test data. These studies often do not integrate temporal information.

[3] Please note that a model trained by a self-supervised learning algorithm is called a self-supervised learning (SSL) model in speech research.

We categorize the most innovative deep neural networks (DNNs) based on their rising popularity and exceptional performance in facial expression recognition (FER) challenges. For example, in Liu et al. (2021), an image enhancement method is first introduced to identify the facial target area and improve image contrast. Following this, a hybrid feature representation technique is presented, combining four different feature extraction methods: Local Binary Patterns, hybrid features, Pyramid Histogram of Oriented Gradients, and Edge Histogram Descriptor. This approach aims to obtain more discriminative features. In Wang et al. (2020a), a region attention network (RAN) is designed to adaptively capture the relevance of face areas for occlusion and pose variation in FER. RAN combines and embeds several region characteristics produced by a backbone CNN into a compact, fixed-length representation, enhancing the network's ability to handle variations in facial expressions. Wang et al. (2020b) improved StarGAN by introducing contextual loss and attention U-Net to fix the faults in the critical sections of the generated face, such as the mouth and the fuzzy side face image. To efficiently integrate information and semantic aspects of images, the attention U-Net is added to StarGAN's generator, replacing StarGAN's original generator. Zhang et al. (2022b) presented an end-to-end model for face image synthesis and FER simultaneously to address appearance differences and insufficient training data. By conducting image generation and representation learning together, these two activities can increase each other's performance. Zhao et al. (2023) suggested a facial landmark encoding approach to construct a graph. They employed a GCN to comprehensively examine the structural information of facial components underlying various expressions. A CNN is also used for the entire face observation to learn the global properties of diverse expressions. These two networks' properties are combined into a comprehensive high-semantic representation that enhances FER reasoning from a visual and structural viewpoint.

Dynamic Image-Based FER Networks. The FER can take advantage of the temporal correlations of subsequent frames in a sequence, whereas most earlier methods rely on static images. This section introduces some dynamic-based FER networks. For example, in Nguyen et al. (2022), a multilevel convolution neural network is applied to each frame for feature extraction then all the features obtained from each frame are concatenated into a single-3-D vector for FER classification. Yu et al. (2020) provide a multitask learning

approach for global-local facial expression mapping. First, a shared shallower module is employed to learn data from local regions and the overall image. Then, using a partbased module, it extracts meaningful local features associated with facial emotions by processing crucial local regions, such as the eyes, nose, and mouth. Wang et al. (2022b) presented a face frontalization using a cascade regression for dynamic facial expression analysis. A cascaded design with three levels of DCNN is employed to increase FER performance. Faces can be predicted in a coarse-to-fine manner by the network. In Uddin et al. (2021), the extraction and modeling of the temporal dynamics of facial emotions from videos were presented, which is a unique method for FER from videos to address challenges of facial expression identification. Nie et al. (2021) introduced a new correlation-based GCN for automated FER that can incorporate the association of intra-class and inter-class videos for feature learning and information fusion in a complete way. This correlated data could aid in the improvement of node feature categorization in the GCN process. In the meantime, the multihead attention is employed to forecast the hidden relationship between the videos, which increases the performance of the classifier by strengthening the inter-class correlation.

2.2 Task Evaluation

The effectiveness of a deep emotional representation is evaluated by performing a classification or regression using these representations as input. For classification, emotion recognition systems use a classification score or accuracy as a metric. However, as data are often imbalanced across the classes, in naturalistic emotion corpora, the accuracy is usually used as the so-called unweighted accuracy (UA) or unweighted average recall (UAR), representing the average recall across classes, unweighted by the number of instances per class. This has been introduced by the first challenge in the field—the Interspeech 2009 Emotion Challenge and has since been used by other challenges in the field. Emotion recognition systems that use deep representation for emotional attributes such as arousal and valence or dominance prediction commonly optimize regression-based models using the mean squared error (MSE) and the concordance correlation coefficient (CCC) as objective functions.

2.3 Datasets

In this section, we introduce the datasets for unimodal emotion recognition as follows:

2.3.1 Text Emotion Recognition Datasets

For text emotion recognition task, we summary the following datasets:

(1) **ISEAR** [4] contains 7,600 self-reported experiments of emotion-provoking text, which are generated by their reactions to seven primary emotions.

(2) **NLPCC-2018 database** [5] contains 7,928 code-switching texts with five emotion labels, and each text contains more than one language (Chinese and English). It was the benchmarking data for NLPCC Shared Task of Emotion Detection.

(3) **EmotionContext** [6] is a collection of three-turn dialogues, and each utterance is annotated with emotion labels (Chatterjee et al., 2019). The training dataset contains 15k records for emotion classes and 15k records of "Others".

(4) **DailyDialog** [7] contains 13,118 multiturn dialogues and 897 scenes of daily life. This dataset is manually annotated with both communication intention and emotion labels.

(5) **EmoryNLP** [8] collects multiparty dialogues from TV show transcripts: "Friends". This dataset comprises 97 episodes, 897 scenes, and 12,606 utterances.

(6) **EmotionLines** (Hsu et al., 2018) contains a total of 29,245 utterances from 2,000 dialogues. They are collected from Friends TV scripts and private Facebook messenger dialogues. The emotions of all utterances are labeled based on the textual content.

(7) **Multimodal EmotionLines Dataset (MELD)** [9] is an expansion of EmotionLines. It contains the same dialogue instances available in EmotionLines, but it also encompasses audio and visual modality along with text.

[4] https://www.unige.ch/cisa/research/materials-and-online-research/research-material/
[5] http://tcci.ccf.org.cn/conference/2018/taskdata.php
[6] https://www.humanizing-ai.com/emocontext.html
[7] http://yanran.li/dailydialog
[8] https://paperswithcode.com/dataset/emorynlp
[9] https://affective-meld.github.io/

(8) **Equity Evaluation Corpus** [10] was the benchmarks for SemEval-2018 Task: Affect in Tweets. 8,640 English sentences are carefully chosen in this dataset to tease out biases towards races and genders. Both 11 emotion or sentiment and their intensity are provided.

(9) **EmoInt Dataset** [11] (also called Tweets-2016) annotates 7,100 English tweets with four emotions: joy, sadness, fear, and anger. The intensity of the emotion is assigned between -1 and 1, indicating the sentiment with -1 being negative and 1 positive. It was the benchmark data for the WASSA-2017 Shared Task on Emotion Intensity.

(10) **SEMAINE Database** [12] contains approximately 240 character conversations. They are the interaction records between a human user and a human operator who pretends to be an artificially intelligent agent with a prototypic emotional character. This study is still ongoing, and currently, approximately 80 conversations have been fully dimensional annotated.

2.3.2 Speech Emotion Recognition Datasets

The speech databases can be broadly classified into three types: Actor (simulated), Elicited (induced) and Natural. Specifically, simulated emotion samples are collected from skilled professionals, such as theater or radio performers, who deliver pre-determined lines in specific moods. To assess expressiveness, recordings can be made in different sessions, accounting for individual differences in speech production. These databases make up more than 60% of the collected emotion databases. Simulated emotions are considered effective for accurately conveying emotions. Elicited emotion samples are created without the speaker's knowledge by placing them in a situation designed to provoke the desired emotions. The speaker is unaware that they are being recorded. Although these databases are more natural than the simulated ones, the speakers might not fully express the intended emotions. Natural emotion samples are collected from real-life situations where emotions are not simulated or elicited. Examples include call center recordings, cockpit recordings,

[10] https://www.saifmohammad.com/WebPages/Biases-SA.html
[11] https://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html
[12] https://ibug.doc.ic.ac.uk/research/semaine-database/

Corpus Name	Language	Speakers	Mode	Type	Emotions	Duration (approx)	Public Access
EMODB [78])	German	10 speakers (5 males, 5 females)	audio	stimulated	anger, boredom, disgust, fear, happiness, sadness, neutral	$<$ 1 hour	yes
MSP-IMPROV [79]	English	12 actors (6 males and 6 females)	audio, video	stimulated	anger, happiness, sadness, neutral	18 hour	yes
MSP-Podcast [80]	English	60 speakers (30 females, 30 males)	audio	naturalistic	arousal, valence, dominance	27 hours	yes
SEMAINE [81]	English	150 participants	audio, video	induced	5 affective dimensions (i. e., valence, activation, power, anticipation/ expectation, intensity)	6.2 hours	yes
IEMOCAP [82]	English	5 females, 5 males	audio, video	stimulated	neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excited and other)	12 hours	yes
EMOVO [83]	Italian	6 speakers (3 males, 3 females)	audio	stimulated	disgust, happiness, fear, anger, surprise, sadness, neutral	$<$ 2 hours	yes
RECOLA [84]	French	46 speakers (19 males, 27 females)	audio, video	natural	five social behaviours (dominance, agreement, performance, engagement, rapport); arousal and valence	3.5 hours	Yes
CMU-MOSEI [85]	English	single speaker	multimodal	natural	anger, anxious, disgust, happiness, neutral, sadness, surprise and fear	65 hours	Yes

Figure 2.1: Review of Different SER Databases (Latif et al., 2023).

and patient-doctor dialogues. These recordings capture genuine emotional expressions. We summarize some important benchmarks as follows:

Different emotional datasets are available; however, in this work, we present only the details of the most popular ones in Table 2.1, which are being used to learn emotional representation, including EMODB (Busso et al., 2017), MSP-IMPROV (Lotfian and Busso, 2019), MSP-Podcast (McKeown et al., 2012), SEMAINE (Busso et al., 2008), IEMOCAP (Costantini et al., 2014), EMOVO (Ringeval et al., 2013), RECOLA (Zadeh et al., 2018b) and CMU-MOSEI (Schuller et al., 2011). In Latif et al. (2023), the authors provide more details on emotional speech databases. These emotional datasets are annotated using categorical or dimensional emotion models.

2.3.3 Facial Emotion Recognition Datasets

For facial emotion recognition task, we summary the following datasets:

(1) **JAFFE.** The Japanese Female Facial Expression (JAFFE) (Lyons et al., 2020) is a freely available dataset, published in 1998 by a group of Japanese researchers from the ATR Human Information Processing Research Laboratory and the Psychology Department of Kyushu University. The dataset features images of 10 Japanese women,

expressing the following emotions: happiness, sadness, surprise, anger, disgust, fear, and neutral. Each participant took around 3 to 4 pictures of herself for each emotion while looking through a semi-reflective plastic sheet toward the camera, totaling 219 images. The camera was also surrounded by a black box to prevent and mitigate light reflections. The hair was removed from the front of the face in a manner that the expressions were more evident. Because the photographic process was performed analogously, the photos were digitized afterward.

(2) **CK.** Cohn-Kanade AU-Coded Expression Database (Kanade et al., 2000) is a well-established dataset, being widely used. The database has two versions. The first, released in 2000, includes 486 sequences from 97 posers, with 69% female, 31% male, 81% Euro-American, 13% Afro-American, and 6% from other groups. It contains sequences from neutral to peak expression, with the peak image fully FACS-coded and given an emotion label based on the requested expression, not necessarily the one actually expressed. Images were taken in a controlled environment with either 640x490 or 640x480 resolution in 8-bit gray-scale or 24-bit color. The second version, the Extended Cohn-Kanade Dataset (CK+), was released in 2010. It includes 107 sequences of emotion transitions from 26 subjects, with non-posed images receiving nominal emotion labels based on the subject's impression of the seven basic emotions: Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise.

(3) **NMI.** The MMI dataset (Mohseni et al., 2014) includes data from 25 individuals of different ethnicities (European, Asian, and South American), with 44% women and 56% men, aged 19 to 62. Expressions were naturally recorded, induced by videos (comedy and disgusting content) to simulate natural conditions. Unique to MMI, it includes both basic and non-basic expressions, capturing expression exchanges. The images are in real color, digitized to 720 x 576 pixels, and videos recorded at 24 fps, with sequences ranging from 40 to 520 frames. The dataset totals 1.5 hour, features varied backgrounds, and is freely available online for the scientific community.

(4) **FER2013.** The FER2013 database (Goodfellow et al., 2015) contains 35,887 grayscale images of faces at 48 x 48 pixels, labeled with seven facial expressions: 4,953 angry, 547 disgust, 5,121 fear, 8,989 happy, 6,077 sad, 4,002 surprise, and 6,198 neutral.

Developed using Google's image search API with 184 emotion-related keywords, the dataset was initially used for a Kaggle competition, providing 28,709 images for training and 3,589 for testing and validation. Post-competition, it was made publicly accessible. Despite its large image count, the dataset has a reduced spatial resolution, suitable for training and testing computational classifiers.

(5) **BU-3DFE.** The Binghamton University 3D Facial Expression (BU-3DFE) dataset (Yin et al., 2006), provides emotion-classified 3D facial images for facial expression recognition research. It includes images linked to emotional states such as anger, disgust, fear, happiness, sadness, surprise, and neutral, with a degree of spontaneity associated with each image. Using a system with six cameras positioned at different angles and stereo photogrammetry, the dataset captures 3D faces. It comprises 2,500 models from 100 participants (60% women, 40% men) from the psychology, arts, and engineering departments, each demonstrating seven emotions at four intensity levels (low, middle, high, and highest). Subsequent related datasets include BU-4DFE, BP4D-Spontaneous, and BP4D+. The BU-3DFE dataset is freely available for use.

2.4 Emotion Recognition Versus Sentiment Analysis

Task	Input (Text)	Input (Visual)	Input (Acoustic)	Target Label
SA	Plot to it than the action scenes were my favorite parts.	···	···	+1.666
ER	Yes, I did notice that. It does look really beautiful over the water.	···	···	Happy

Table 2.1: Example of data formats for SA and emotion recognition tasks. The target "$+1.666$" for SA task indicates a positive sentiment strength of 1.666, while the target "Happy" for emotion recognition task indicates an emotion category of happy.

To provide a detailed differences between Sentiment Analysis (SA) and Emotion Recognition (ER), we list three core differences as follows:

(1) **Formulated Definitions:** SA aims to predict the real number $y_i \in \mathbb{R}$ that reflects the sentiment strength. However, emotion recognition aims to predict the emotion category, $y_i \in \{c_1, \cdots, c_k\}$ with k emotion categories, for each utterance (Joshi et al., 2022). An example of data formats for two tasks is listed in Table 2.1. We also provide an example for

the SA task as:
[input, label] = [input = (t = Plot to it than the action scenes were my favorite parts. $a = \ldots$, $v = \ldots$), label = +1.666] where the label $y_i = +1.666$ means the positive strength is 1.666 and the SA task is a regression task.
An example for the emotion recognition task as:
[input, label] = [input = (t = Yes, I did notice that. It does look really beautiful over the water. $a = \ldots$, $v = \ldots$), label=Happy] where the label is Happy and emotion recognition is a classification task.
(2) **Loss Function:** for the SA task, since it is a regression task, BBFN (Han et al., 2021a) used mean squared error (MSE) loss and MMIM (Han et al., 2021b) used mean absolute error (MAE) loss for the regression task. MSE loss in BBFN is formulated as:

$$\mathcal{L}_{task} = \text{MSE}(\hat{y}_i, y_i) = \frac{1}{n}\sum_{i=1}^{n} \|\hat{y}_i - y_i\|_2^2, \tag{2.1}$$

where $\hat{y}_i$ is the predicted value for the i-th sample, y_i is the truth label for the i-th label, n is the total number of samples and $\|\cdot\|_2$ is the L_2 norm. And MAE loss is formulated as:

$$\mathcal{L}_{task} = \text{MAE}(\hat{y}_i, y_i) = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|, \tag{2.2}$$

where $|\cdot|$ is the L_1 norm.

For emotion recognition task, same as other emotion recognition work (Hu et al., 2022a), we adopt cross-entropy loss as:

$$\mathcal{L}_{task} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_i^j \log(\hat{y}_i^j), \tag{2.3}$$

where k is the number of emotion classes, n is the number of utterances, $\hat{y}_i^j$ represents the predicted probability that the i-th sample belongs to the j-th class, and y_i^j is the ground truth for the i-th class belongs to the j-th class.

Through different task-guided loss function, SA methods can well capture the features that well characterize the strength (+1.66 or -1.66) of positive or negative, while emotion

recognition methods focus on capturing the features that well characterize the detailed emotion (happy or sad) of each utterance. Thus, our model can well classify emotion classes but may not have a good prediction for the strength of sentiments.

(3) **Metrics**: For the MSE task, current methods (Han et al., 2021a,b; Yu et al., 2021) almost adopt mean absolute error (MAE), Pearson correlation (Corr), two-class classification accuracy (ACC-2), seven-class classification accuracy (ACC-7) and F1 score computed for positive/negative and non-negative/negative classification as evaluation metrics.

For the emotion recognition task, current methods (Joshi et al., 2022; Wei et al., 2019; Hu et al., 2022a) adopt F1 score to measure each emotion, Accuracy for emotion recognition and weighted F1 (WF1) for evaluation. Although the MSE task and emotion recognition task all adopt Accuracy metric, the MSE task is under a more relaxed condition (Han et al., 2021a,b; Yu et al., 2021). Specifically, for the binary classification task on MOSEI, the MSE methods can treat all predicted value $\hat{y}_i \leq 0$ as negative class and all predicted value $\hat{y}_i$ ¿ 0 as positive class. However, for the emotion recognition task, the predicted label $\hat{y}_i \in \{\text{positive}, \text{negative}\}$, which is under a more severe conditions (Joshi et al., 2022; Wei et al., 2019; Hu et al., 2022a). So MOSEI methods such as COGMEN (Joshi et al., 2022) do not compare with MSE methods even on the same metrics of ACC-2 and ACC-7.

Chapter 3

An Empirical Study: Speech Emotion Recognition

In this chapter, we will provide an empirical study on speech emotion recognition. Please note that our proposed models can be applied simply to other unimodal emotion recognition tasks such as video, text, and image.

Before delving into the details of our proposed methods, we give a brief introduction to current speech emotion recognition (SER) models and their limitations in Section 3.1. Specifically, existing methods often overlook the information gap between the pre-training speech recognition task and the downstream speech emotion recognition task, resulting in sub-optimal performance. Moreover, current methods require much time for fine-tuning, which limits their effectiveness in real-world scenarios with large-scale noisy data.

Then, we introduce the necessary background knowledge and related work in Section 3.2. We highlight the main differences of our proposed method with existing methods.

Next, in Section 3.3, we propose an active learning (AL)-based fine-tuning framework for SER, called AFTER, that leverages task adaptation pre-training (TAPT) and AL methods to enhance performance and efficiency. Specifically, we first use TAPT to minimize the information gap between the pre-training speech recognition task and the downstream SER task. Then, AL methods are employed to iteratively select a subset of the most informative and diverse samples for fine-tuning, thereby reducing time consumption.

We present the empirical results of our model on the IEMOCAP and three constructed

MERGED datasets in Section 3.4 and Section 3.5. Experiments demonstrate that our proposed method AFTER, using only 20% samples, improves precision by 8. 45% and reduces time consumption by 79%. The additional extension of AFTER and ablation studies further confirms its effectiveness and applicability to various real-world scenarios.

Finally, we give the limitations and summaries of this study in Section 3.6.

3.1 Introduction

3.1.1 Current Studies of Speech Emotion Recognition

"The language of tones is the oldest and most universal of all our means of communication" (Blanton, 1915). Speech emotion recognition aims to identify emotional states conveyed in vocal expressions, making it an essential topic in tone and language analysis. It has gained significant attraction in both the industrial and academic communities, including speech-to-text translation (Taghavi et al., 2023; Santoso et al., 2022), dialogue system (Wang et al., 2024b; Li et al., 2023b), medical surveillance systems (Clavel et al., 2008), psychological treatments (Elsayed et al., 2022; Li et al., 2022a), and intelligent virtual voice assistants (Wang et al., 2023).

With the development of deep learning techniques in natural language processing (Zhang et al., 2021b, 2024) and computer vision (Wang et al., 2024a), many speech emotion recognition methods have been proposed. These methods are broadly classified into machine learning-based methods and deep learning-based methods (BJ.Abbaschian et al., 2021).

Specifically, machine learning-based methods (Gharsellaoui et al., 2019; Paraskevopoulos et al., 2019) typically consist of three main components: feature extraction, feature selection, and emotion recognition. However, selecting and designing features for specific corpora is time-consuming (Ayadi et al., 2011), and they consistently exhibit poor generalization on unseen datasets (Padi et al., 2021). Deep learning-based methods can address these issues by automatically extracting more abstract features to improve generalization (LeCun et al., 2015; Morais et al., 2022; Chen et al., 2023a). They benefit from various neural network architectures, such as convolutional neural networks (CNNs) (Aftab et al., 2022) and transformers (Ghriss et al., 2022). With the development of pre-trained

language models (Devlin et al., 2019) and the availability of large-scale datasets, various pre-trained automatic speech recognition models have been proposed. [1] These automatic speech recognition models use speech's acoustic and linguistic properties to provide more robust and context-aware representations for speech signals. Xia et al. (2021) proved that fine-tuning speech emotion recognition datasets on wav2vec 2.0 (Schneider et al., 2019) obtains state-of-the-art performance on IEMOCAP (Busso et al., 2008). This finding has inspired researchers to explore new fine-tuning strategies in automatic speech recognition models, becoming a new paradigm for speech emotion recognition. For example, Ren et al. (2022) proposed a self-distillation speech emotion recognition model to fine-tune wav2vec 2.0 obtaining state-of-the-art performance on the DEMoS dataset. And Ferreira (2022) fine-tuned wav2vec 2.0 by jointly optimizing speech emotion recognition and automatic speech recognition tasks, achieving state-of-the-art performance in Portuguese datasets.

3.1.2 Limitations of Previous Studies

Although the aforementioned methods achieve considerable success, several issues still need to be addressed.

(1) Current methods seldom consider the information gap between the pre-trained automatic speech recognition and downstream speech emotion recognition tasks. For example, wav2vec 2.0 (Baevski et al., 2020) adopts a masked learning objective to predict missing frames from the remaining context, while the downstream speech emotion recognition (Aftab et al., 2022; Baruah and Banerjee, 2022) task aims to minimize cross-entropy loss between predicted and referenced emotion labels for speech signals. Suchin et al. (Gururangan et al., 2020) proved that the information gap would decrease the performance of downstream tasks. To address it, Pseudo-TAPT (L.Chen and A.Rudnicky, 2022) first uses K-means to obtain pseudo-labels of speech signals and uses supervised TAPT (Gururangan et al., 2020) for continual pre-training. However, K-means is sensitive to the initial value, making Pseudo-TAPT unstable and computationally expensive.

[1] In this draft, automatic speech recognition models refer to those models that use machine learning or artificial intelligence technology to process human speech into readable text such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Babu et al., 2021) and Data2vec (Baevski et al., 2022).

(2) Current methods only fine-tune and validate the performance on a specific speech dataset. For example, Xia et al. (2021) train their models solely on the IEMOCAP, leading to over-fitting and poor generalization for unseen datasets. Real-world scenarios contain much heterogeneous and noisy data, which hinders the application of these speech emotion recognition methods. Heterogeneous means that real-world scenarios should contain different voice background, languages, devices for recording speech, and speech types (spontaneous speech and acted speech). Please note that "noisy data" does not refer to acoustically noisy data (unclear speech or unrecognized audio). We define "noisy data" as the specific noise in the speech emotion recognition task, including outliers and redundant samples. Specifically, outliers encompass various ambiguous emotions due to the complexity of speech, which can lead to inaccurate emotional annotations and degrade the performance of the model. Redundant samples being trained repeatedly does not improve the model's accuracy. Instead, they lead to an uneven distribution of data, making it more challenging to identify emotions with a limited amount of data.

(3) Pre-trained automatic speech recognition models often contain millions of parameters, for example, wav2vec 2.0 contains 317 million parameters, which is time-consumption for real-world and large-scale datasets.

3.1.3 Proposed Solutions

To address the aforementioned issues, we propose an active learning-based fine-tuning framework for speech emotion recognition, referred to as AFTER, which can be easily applied to noisy and heterogeneous real-world scenarios. Specifically, we first propose an unsupervised task adaptation pre-training (TAPT) method (Gururangan et al., 2020) to reduce the information gap between the pre-trained and downstream speech emotion recognition tasks, enabling the pre-trained model to understand the semantic information of the speech emotion recognition task. Then, we create two large-scale heterogeneous and noisy datasets to simulate real-world scenes. Furthermore, we propose AL strategies with clustering-based initialization to iteratively select a smaller, more informative, and diverse subset of samples for fine-tuning. This approach can efficiently eliminate noise and outliers,

improve generalization, and reduce time consumption.

3.2 Related Work and Background Knowledge

In this section, we would like to introduce the most related works and background knowledge of our proposed method. Since the main components of our proposed frameworks include active learning and task adaptation pre-training, we will introduce the active learning in Section 3.2.1 and task adaptation pre-training in Section 3.2.2.

3.2.1 Active Learning

Active learning is an extensively research challenge in the field of machine learning, encompassing a variety of scenarios and query strategies (Xu et al., 2013; Zhang et al., 2022d; Li et al., 2024). In recent years, there has been a resurgence of interest in active learning within the NLP community (Zhang et al., 2022d). Recent studies have used active learning with BERT models for specific tasks such as intent classification (Zhang and Zhang, 2020), sentence matching (Bai et al., 2020), parts-of-speech tagging (Chaudhary et al., 2021), and named entity recognition (Liu et al., 2022). Margatina et al. (2021) advocate for continued pre-training on unlabeled data in the context of active learning. Rotman and Reichart (2022) adapt active learning for multi-task scenarios involving transformer models. Ein-Dor et al. (2020) conduct an extensive empirical study of existing active learning strategies in binary classification tasks. Yuan et al. (2020) adapt the BADGE (Ash et al., 2020) framework for active learning with BERT. However, BADGE computes gradient embeddings from the output layer of a neural network and subsequently clusters the gradient space. To the best of our knowledge, our work is the first active learning-based fine-tuning framework in the speech domain. Instead of focusing on proposing complex active learning query strategies, we concentrate on evaluating the effectiveness of active learning for SER. Through experimentation, in Section 3.5.2, we validate its effectiveness and aspire to propose more efficient methods to advance this task in the future.

3.2.2 Task Adaptation Pre-training

Task-adaptive pre-training (TAPT) is a significant area of research, as introduced by (Gururangan et al., 2020). Essentially, TAPT involves customizing a language model for a specific task, leading to improved model performance. Gururangan et al. (2020) explore the benefits of tailoring a pre-trained model like RoBERTa to a specific task domain. They investigate four distinct domains, covering biomedical and computer science publications, news and reviews, and spanning eight classification tasks. Their exploration expanded to assessing the transferability of adapted language models across different tasks and domains. Additionally, they conduct a study to evaluate the importance of pre-training on human-curated data. Konle and Jannidis (2020) discuss various strategies for adapting BERT and DistilBERT to historical domains and tasks in computational humanities. The outcomes support the idea of continuous pre-training in machine learning tasks to enhance performance stability. A combined approach of domain adaptation and task adaptation shows positive effects. Task adaptation alone is versatile and applicable in various setups, unlike domain adaptation, which requires a substantial amount of in-domain data. Several approaches have been explored to make TAPT more efficient, especially with methods involving word embeddings. For example, Nishida et al. (2021) propose TAPTER, enhancing pre-trained language model embeddings for domain adaptation. It outperforms standard methods when in-domain data is limited. El Boukkouri (2021) advocate re-training from a general model for low-resource scenarios, yielding comparable performance with slight trade-offs. Sachidananda et al. (2021) adapt tokenizers to transfer pre-trained models to new domains, achieving 97% performance benefits but introducing a 6% increase in model parameters. In this study, we introduce a straightforward approach for continuous training of a pre-trained model with a task-related loss function on downstream tasks. Our experimental results, detailed in Section 3.5.3, demonstrate the effectiveness of this method.

3.3 Methodology: AFTER

The overall framework of Active Learning-based fine-tuning with task adaptation pre-trining for SER, called AFTER, is depicted in Figure 3.1, comprising three main components: a *task*

adaptation pre-trained module, an *active learning-based fine-tuning* module, and an *emotion classification* module. First, we will formally define the task of SER, and subsequently introduce each component of AFTER in detail.

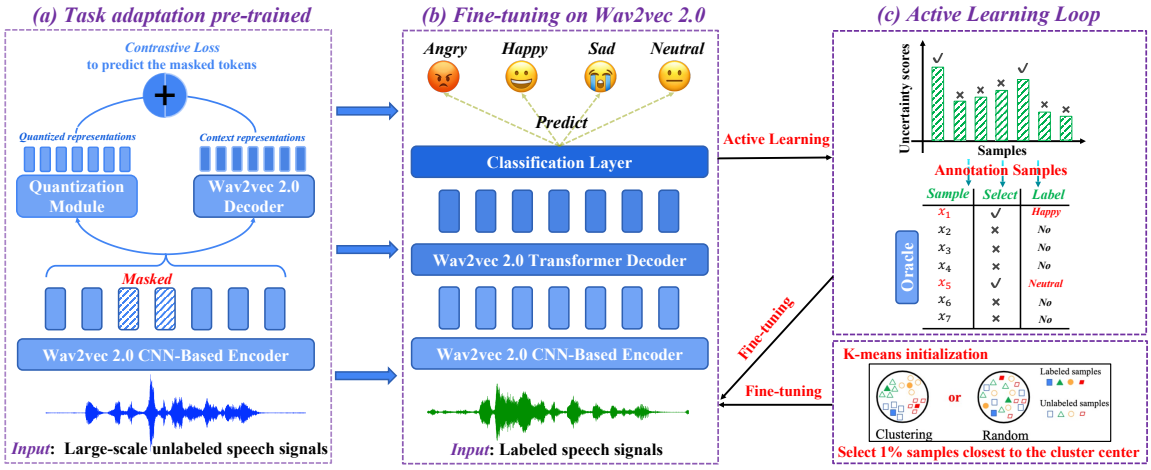


Figure 3.1: Model overview. First, we pre-train an off-the-shelf wav2vec 2.0 in the TAPT manner. Then, we adopt an active learning method to select unlabeled samples for iterative annotation. These labeled samples are used to fine-tune the wav2vec 2.0 model for SER.

3.3.1 Notations and Task Formulation

Given a speech dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where $\mathbf{x}_i$ represents the i-th speech signal and y_i represents its corresponding emotion label, we aim to fine-tune a pre-trained automatic speech recognition model M, such as wav2vec 2.0, on the labeled speech datasets $\mathcal{D}_{\text{train}}$ to obtain accurately predicted emotion labels for all speech signals.

3.3.2 Task Adaptation Pre-training

As shown in Figure 3.1 (a), we introduce the TAPT component in detail. To better leverage pre-trained prior knowledge for the benefit of downstream tasks and minimize the information gap between pre-trained tasks and downstream tasks, Gururangan et al. (2020) continued training the pre-trained model RoBERTa (Liu et al., 2019) on downstream datasets via the same loss of the pre-training task (reconstructing the masked tokens of input sentences,

similar to BERT (Devlin et al., 2019)), resulting in significant improvements in downstream text classification tasks. Inspired by their work, we added an additional step into AFTER by continuing training the pre-trained automatic speech recognition model on downstream training datasets for the speech emotion recognition task. By conducting this process, we can bridge the information gap between the pre-trained automatic speech recognition task and the target SER task, as confirmed by our experiments in section 3.5.3.

As depicted in Figure 3.1 (a), the wav2vec 2.0 model $\mathbf{M}(\mathbf{W}_0)$, with pre-trained weights $\mathbf{W}_0$, consists of three sub-modules: the feature encoder module, the transformer module, and the quantization module. Specifically, we utilize a CNN-based encoder to encode the i-th input unlabeled speech signals into low-dimensional vectors, denoted as $\mathbf{x}_i$. Subsequently, we randomly mask 15% of the features (following BERT (Devlin et al., 2019)) of the speech vectors. We then decode them using two decoders to obtain quantized and context representations. The quantization decoder can transform continuous speech vectors $\mathbf{x}_i$ into discrete codewords from phonemes codebooks [2], resulting in $\mathbf{z}_i^q$. Meanwhile the wav2vec 2.0 decoder (transformer layers) employs self-attention to decode continuous speech vectors $\mathbf{x}_i$ into context-aware representations $\mathbf{z}_i^c$. Then, we design contrastive loss (cl) to minimize the differences between quantized and context representations as follows:

$$\mathcal{L}_{cl} = -\sum_{i=1}^{n} \log \frac{\exp(\text{sim}(\mathbf{z}_i^c, \mathbf{z}_i^q)/\kappa)}{\sum_{j=1}^{n} \exp(\text{sim}(\mathbf{z}_i^c, \mathbf{z}_j^q)/\kappa)}, \tag{3.1}$$

where the temperature hyperparameter κ is set to 0.1, and $\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}/\|\mathbf{a}\|_2\|\mathbf{b}\|_2$, where T represents the transposition of a vector. Eq. (3.1) can help obtain better quantized and context representations because two decoders can provide highly heterogeneous contexts for each speech signal (You et al., 2020).

To minimize the information gap between the pre-trained model and downstream SER task, following BERT (Devlin et al., 2019), we first randomly mask 15% of the tokens of each speech signal. We then apply reconstruction loss on the corrupted downstream SER

[2] A quantized codebook refers to a set of predetermined values or codewords used to represent a continuous signal in a discrete form (Baevski et al., 2020).

dataset to generate tokens for reconstructing the original data, formulated as follows:

$$\mathcal{L}_{rl} = -\frac{1}{|N_m|} \sum_{i=\text{First masked token}}^{\text{Last masked token}} s_i^{\text{true}} \log(s_i^{\text{predicted}}) \tag{3.2}$$

where N_m is the number of masked tokens, s_i^{true} and $s_i^{\text{predicted}}$ are the ground-truth and predicted token probability of the i-th masked token, respectively.

Finally, we combine contrastive loss and reconstruction loss for the TAPT process as:

$$\mathcal{L}_{TAPT} = \mathcal{L}_{cl} + \mathcal{L}_{rl}. \tag{3.3}$$

Please note that although pseudo-TAPT (L.Chen and A.Rudnicky, 2022) also adopts TAPT, we employ different loss functions. We believe that our method is simpler and more suitable for upstream automatic speech recognition tasks. Specifically, they invest significant time using K-means to extract frame-level emotional pseudo-labels and continually pre-train their model in a supervised manner by predicting their frame-level emotion pseudo-labels. However, K-means is sensitive to the initial value and outliers (Zhang et al., 2020), making Pseudo-TAPT unstable and computationally expensive.

Algorithm 1: Active Learning based Fine-tuning

Input : Unlabeled data $\mathcal{D}_{\text{pool}}$, Model $\mathbf{M}(\mathbf{W}_0)$, Acquisition size k, Iterations τ, total number of selected samples N_c, and Acquisition function ac().

$\mathbf{M}_{\text{TAPT}}\ (\mathcal{D}_{\text{pool}}; \mathbf{W}_0') \leftarrow$ Train $\mathbf{M}(\mathbf{W}_0)$ on $\mathcal{D}_{\text{pool}}$;
$Q_0 \leftarrow$ Clustering-based initialization from $\mathcal{D}_{\text{pool}}$;
$\mathcal{D}^0_{\text{train}} \leftarrow Q_0$; $\mathcal{D}^0_{\text{pool}} \leftarrow \mathcal{D}_{\text{pool}} \backslash Q_0$ where $|Q_0| = 1\% N_s$;
$\mathbf{M}_0([\mathbf{W}_0', \mathbf{W}_c]) \leftarrow$ Initialized from $\mathbf{M}_{\text{TAPT}}\ (\mathcal{D}_{\text{pool}}; \mathbf{W}_0')$;
$\mathbf{M}_0(\mathcal{D}^0_{\text{train}}; [\mathbf{W}_0', \mathbf{W}_c]) \leftarrow$ Train $\mathbf{M}_0([\mathbf{W}_0', \mathbf{W}_c])$ on $\mathcal{D}^0_{\text{train}}$;
for *i* $\leftarrow$ *1 to* τ **do**
    $Q_i \leftarrow \text{ac}(\mathbf{M}_{i-1}, \mathcal{D}^{i-1}_{\text{pool}}, k)$ ▷ Annotating k samples;
    $\mathcal{D}^i_{\text{train}} = \mathcal{D}^{i-1}_{\text{train}} \cup Q_i$ ▷ Add labeled samples to $\mathcal{D}^i_{\text{train}}$;
    $\mathcal{D}^i_{\text{pool}} \leftarrow \mathcal{D}^{i-1}_{\text{pool}} \backslash Q_i$ ▷ Delete samples from $\mathcal{D}^i_{\text{pool}}$;
    $\mathbf{M}_i(\mathcal{D}^i_{\text{train}}; [\mathbf{W}_0', \mathbf{W}_c]) \leftarrow$ Train $\mathbf{M}_{i-1}$ on $\mathcal{D}^i_{\text{train}}$;
end

3.3.3 Active Learning based Fine-tuning

When we complete the TAPT process, we obtain the model $\mathrm{M}_{\mathrm{TAPT}}(\mathbf{W}_0^{'})$ with $\mathbf{W}_0^{'}$ as the initialization of the weight for the active learning process (cf. Line 1 of Algorithm 1). A typical active learning setup starts by treating $\mathcal{D}$ as a pool of unlabeled data $\mathcal{D}_{\text{pool}}$ and performs τ iterations of sample selection. Specifically, in the i-th iteration, k samples are selected using a given acquisition function ac(): $Q_i = \{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$. Here, k is a variable parameter. We determine it based on the number of iterations τ and the predefined total number of selected samples N_s, i.e., $k = \text{ROUND}(N_s/\tau)$, where ROUND() rounds the number down. For example, we adopt Entropy (R.Nicholas and M.Andrew, 2001) as the ac() function to measure the uncertainty of the samples and select the most uncertain k samples. These selected samples are then labeled and added to the i-th training dataset $\mathcal{D}_{\text{train}}^{i}$, with which a model is fine-tuned for SER.

One primary goal of AFTER is to explore whether active learning strategies can reduce the number of annotation samples, as labeling large-scale datasets is the most laborious part of SER. Instead of focusing on proposing new active learning query strategies, we adopt five of the most well-known and influential active learning strategies for evaluation, including Entropy, Least Confidence (M.Dredze and K.Crammer, 2008), Margin Confidence (M.Dredze and K.Crammer, 2008), ALPs (M.Yuan et al., 2020), and BatchBald (A.Kirsch et al., 2019). These methods use different criteria to help select the most uncertain and informative samples from $\mathcal{D}_{\text{pool}}$, and we introduce them briefly as follows:

(1) Entropy measures the uncertainty of $\mathbf{x}_i$ as

$$\text{Entropy}(\mathbf{x}_i) = -\sum_{j=1}^{c} P(\hat{y}_j|\mathbf{x}_i)\text{log}P(\hat{y}_j|\mathbf{x}_i), \tag{3.4}$$

where c is the number of emotional classes and $P(\hat{y}_j|\mathbf{x}_i)$ represents the predicted probability of $\mathbf{x}_i$ for the j-th emotion.

(2) Least Confident measures the most incontinent samples as

$$\text{Least Confident}(\mathbf{x}_i) = \sum_{j=1}^{c}(1 - P(\hat{y}_j|\mathbf{x}_i)) \tag{3.5}$$

where c is the number of emotional classes and $P(y_j|\mathbf{x}_i)$ represents the predicted probability of $\mathbf{x}_i$ for the j-th emotion.

(3) Margin Confidence is the process of selecting the sample with the smallest difference between the maximum and second largest probability predicted by the model, which can be formulated as

$$\text{Margin Confidence}(\mathbf{x}_i) = (P(\hat{y}_1|\mathbf{x}_i) - P(\hat{y}_2|\mathbf{x}_i)) \tag{3.6}$$

where $P(y_1|\mathbf{x}_i)$ represents the largest predicted probability of $\mathbf{x}_i$ and $P(y_2|\mathbf{x}_i)$ represents the second largest predicted probability of $\mathbf{x}_i$.

(4) ALPs iteratively selects the sample closest to the cluster center as the most differentiated and informative sample each batch. This can be formulated as follows:

$$\text{ALPs}(\mathbf{x}) = \text{argmin}_{\mathbf{x}_i} \|\textbf{centers} - \mathbf{x}_i\| \tag{3.7}$$

where **centers** is the clustering centers (we follow their paper using K-means to obtain clustering centers).

(5) BatchBald jointly score samples by estimating the mutual information ($\mathbb{I}$) between a set of multiple data points and the model parameters:

$$\text{BatchBald}(\{\mathbf{x}_1, \ldots, \mathbf{x}_b\}, p(\mathbf{w}|\mathcal{D}_{\text{train}})) = \mathbb{I}(y_1, \ldots, y_b; \mathbf{w}|\mathbf{x}_1, \ldots, \mathbf{x}_b, \mathcal{D}_{\text{train}}). \tag{3.8}$$

After applying the above query strategies for the samples, we select the most uncertain or diverse k samples for annotation and add them to the training dataset $\mathcal{D}_{\text{train}}$. Traditional active learning methods often use random initialization; however, we found that these active learning methods are sensitive to the initialization process, leading to the selection of redundant samples or outliers in each active learning iteration with poor initialization. Therefore, instead of directly using active learning methods, we propose a clustering-based initialization for all active learning methods (K-means in this study), resulting in better performance. Please note that, as illustrated in Algorithm 1, clustering-based initialization

is applied only in the initialization process, and subsequent iterations of the active learning loop do not require a K-means process.

3.3.4 Emotion Recognition Classifier

As shown in Figure. 3.1 (b), we incorporate a task-specific classification layer with additional parameters $\mathbf{W}_c$ for emotion recognition on top of wav2vec 2.0. We fine-tune the classification model $\mathbf{M}_i([\mathbf{W}_0^{'}, \mathbf{W}_c])$ in each active learning iteration using all labeled samples in $\mathcal{D}_{\text{train}}$ (cf. Lines 6-10 of Algorithm 1). We formulate the cross-entropy loss for the emotion recognition classifier as follows:

$$\mathcal{L}_{ce} = -\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{c} y_i^j \log(\hat{y}_i^j), \tag{3.9}$$

where c is the number of emotion classes, k is the number of selected samples at t-th iteration, $\hat{y}_i^j$ is the i-th predicted label, and y_i^j is the i-th ground-truth of the j-th class.

3.4 Experimental Settings

In this section, we first introduce all datasets used in this study in Section 3.4.1. Following this, we present the selected baselines in Section 3.4.2 and provide implementation details in Section 3.4.3. Finally, we delve into the detailed active learning strategies used in the following experiments in Section 3.4.4.

3.4.1 Datasets

IEMOCAP: We first evaluated the performance of all baseline models using the widely used benchmark dataset, IEMOCAP (Busso et al., 2008). IEMOCAP is a multimodal database commonly used to evaluate SER performance. It comprises five conversation sessions, each featuring a female and a male actor engaging in improvised and scripted scenarios. The dataset includes 10,039 speech utterances, all sampled at 16kHz with a 16-bit resolution. To ensure a fair comparison, we merged the "excited" class into the "happy"

class, resulting in four considered emotions: neutral, happy, angry, and sad. Following the approach of L.Chen and A.Rudnicky (2022), we adopted a five-fold cross-validation method, where each IEMOCAP session served as the test set. Furthermore, we randomly selected 10% of the data from the remaining four sessions for our validation dataset, with the rest allocated to our training dataset.

SAVEE: To explore the performance of AFTER with broader range of emotions, we incorporated an additional datasets, the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset (Jackson and Haq, 2014). SAVEE contains four male speakers: DC, JE, JK, and KL. Each speaker reads the same set of 120 sentences, labeled with one of seven emotion categories: angry, disgust, sad, fear, happy, surprise, and neutral. Utilizing all emotion categories, the dataset comprises 480 utterances, totaling 30 minutes of speech. For fair comparisons with SOTA approaches in experiments, following the previous works (Tuncer et al., 2021; Ye et al., 2022; Wen et al., 2022; Farooq et al., 2020; Ye et al., 2023), we mainly conducted 10- fold cross-validation. In each fold, we allocated 90% of the data for training and 10% for testing to evaluate the model's fitting ability.

Merged-I dataset: Many existing methods are inadequate for real-world applications and are susceptible to noise due to their heavy reliance on fine-tuning models using specific small-scale datasets. For example, pseudo-TAPT (L.Chen and A.Rudnicky, 2022) is fine-tuned by the training dataset of IEMOCAP and performs well on IEMOCAP. However, pseudo-TAPT performs poorly when tested on other datasets. To provide a potential solution to address this issue, we conducted additional experiments by training on two larger noisy and heterogeneous datasets. We achieved this by merging various datasets from different sources to simulate the noisy environments encountered in real-world scenarios. It is important to note that any emotion recognition datasets containing the corresponding emotional categories can be incorporated into the Merged-I datasets. In this draft, we selected five widely-used datasets with different languages, recording equipment, and number of actors. We first introduce each dataset of the Merged-I dataset as follows:

- EmoDB (Burkhardt et al., 2005) database is a freely available German emotional database, created by the Technical University. It features ten professional speakers,

including five males and five females, who participated in the data recording process. The database contains a total of 535 utterances and comprises seven emotions: anger, boredom, anxiety, happiness, sadness, disgust, and neutral. The data was recorded at a 48-kHz sampling rate and then down-sampled to 16-kHz.

- ShEMO (OM.Nezami, 2019) database includes 3,000 semi-natural utterances, totaling three hours and 25 minutes of speech data extracted from online radio plays. ShEMO encompasses speech samples of 87 native-Persian speakers, covering six basic emotions: anger, fear, happiness, sadness, surprise, and neutral states.

- RAVDESS (SR.Livingstone and FA.Russo, 2018) database contains 7,356 files and features 24 professional actors (12 female, 12 male). Each actor vocalizes two lexical-matched statements in a neutral North American accent. The speech includes calm, happy, sad, angry, fearful, surprise, and disgust. Each expression is produced at normal and strong, alongside an additional neutral expression. All conditions are available in three modality formats: Audio-only, Audio-Video, and Video-only.

- EMov-DB (A.Adaeze et al., 2018) includes recordings from four speakers, including two males and two females. The emotional styles covered include neutral, sleepiness, anger, disgust, and amused. Each audio file is recorded in 16bits .wav format.

- CREMA-D (Cao et al., 2014) database is an emotional multimodal actor dataset consisting of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors ranging in age from 20 to 74, representing a variety of races and ethnicities. Actors speak from a selection of 12 sentences, each presented with one of six different emotions: anger, disgust, fear, happy, neutral, and sad, and across four different emotion levels: low, medium, high, and unspecified.

As shown in Table 3.1, we manually controlled the number of instances for each of the four labels in the Merged-I dataset to maintain the balance of the labels. Different from IEMOCAP, EmoDB is a German emotional database, ShEMO is a Persian emotional database, and both RAVDESS and CREMA-D contain more actors (24 actors and 91 actors, respectively). We constructed the Merged-I dataset by merging the training data of the following mentioned datasets with the training data of IEMOCAP. To explore whether the

Table 3.1: Descriptive statistics of the Merged-I dataset, which contains speeches in three types of languages. The ratio of the four labels is Anger : Neutral : Sad : Happy.

Datasets	***Characteristics***		
	# Samples	# Actors	Ratio of Four Labels
IEMOCAP (Busso et al., 2008) (English)	10,038	2	2.5 : 1.2 : 2.4 : 1.0
EmoDB (Burkhardt et al., 2005) (German)	408	10	3.1 : 1.3 : 1.0 : 1.1
ShEMO (OM.Nezami, 2019) (Persian)	2,737	87	5.3 : 5.1 : 2.2 : 1.0
RAVDESS (SR.Livingstone and FA.Russo, 2018) (English)	672	24	2.0 : 1.0 : 2.0 : 2.0
EMov-DB (A.Adaeze et al., 2018) (English)	3,038	4	1.4 : 1.0 : 0.0 : 0.0
CREMA-D (Cao et al., 2014) (English)	4,900	91	1.0 : 1.7 : 1.0 : 1.0
Merged-I dataset	21,793	218	1.5 : 1.4 : 1.0 : 1.5

Merged-I dataset could improve performance on a single dataset, such as IEMOCAP, we employed a 5-fold cross-validation approach. This involved leaving each IEMOCAP session out as the test set and randomly selecting 10% of the dataset from the remaining Merged-I dataset as our validation dataset, while the remainder was allocated for training purposes. It is important to note that we only use the training data for the TAPT and AL-based fine-tuning processes to prevent data leakage during the evaluation. Furthermore, training procedures are conducted from scratch separately for the IEMOCAP, SAVEE, Merged, Merged-II, and Merged-III datasets.

Merged-II dataset: Merged-I dataset contains acted speech datasets. To better simulate "real-world" scenarios, we incorporated two additional spontaneous datasets to Merged-I dataset to construct Merged-II dataset: The two added datasets are as follows:

- AFEW5.0 (Dhall et al., 2015) is a spontaneous audiovisual emotional video dataset developed for emotion recognition in the wild (EmotiW) challenge in 2015. It contains seven emotional categories: anger, disgust, fear, joy, neutral, sadness, and surprise. These emotions were annotated by 3 annotators. The dataset is divided into three parts: train set (723 samples), validation (val) set (383 samples), and test set (539 samples). In this work, we used the train and val sets to validate the performance of our method since the Test set is only available to participants in competitions.

- BAUM-1s (Zhalehpour et al., 2017) is a spontaneous audiovisual affective face database of affective and mental states developed in 2016, featuring 31 Turkish individuals. The video samples were collected in real scenarios, where emotions were elicited by watching films in an unscripted and unguided way. The target emotions include seven basic ones: joy, anger, sadness, disgust, fear, neutral, and surprise, as well as boredom and contempt. Several mental states, such as unsure (confused, and undecided), thinking, concentrating, and bothered, are also included.

We only used four emotion categories from AFEW5.0 and BAUM-1s, including anger, neutral, sad, and happy, to construct the Merged-II dataset. And we used the test data of IEMOCAP as test data for the Merged-II dataset. Training and evaluation processes are similar to those of the Merged-I dataset.

Merged-III dataset: To demonstrate the performance of AFTER on spontaneous datasets, we used the test set of BAUM-1s, which comprises seven emotion categories, for evaluation. The detailed implantation for evaluation follows the approach outlined in (Zhang et al., 2022a). We constructed the Merged-III dataset by merging two acted speech datasets (EmoDB, RAVDESS) with two spontaneous datasets (AFEW5.0, BAUM-1s).

3.4.2 Baselines

We selected different SOTA baselines for different datasets.

(1). For the IEMOCAP dataset, Merged-I dataset, and Merged-II dataset, we selected the best-performing methods: LSSED (Fan et al., 2021), GLAM (Zhu and Li, 2022), RH-emo (Guizzo et al., 2022), Light (Aftab et al., 2022), Pseudo-TAPT (L.Chen and A.Rudnicky, 2022), and w2v2-L-robust (Wagner et al., 2023).

(2). For the SAVEE dataset, we selected the recently best-performing approaches: DCNN (Farooq et al., 2020), TSP+INCA (Tuncer et al., 2021), CPAC (Wen et al., 2022), GM-TCN (Ye et al., 2022), and TIM-Net (Ye et al., 2023).

(3). For the Merged-III dataset, we select the baselines as MFCC+PLP+SVM (Zhalehpour et al., 2017), CNN+SVM (Zhang et al., 2018b), CNN+DTPM+SVM (Zhang et al.,

2018a), CNN+SVM (Ma et al., 2019) and CNN+LSTM (Zhang et al., 2022a).

3.4.3 Implementation details

All experiments used the same learning rate of 10^{-4} with the Adam optimizer. Our implementation of wav2vec 2.0 (wav2vec2-base) is based on the Hugging Face framework [3]. The audio window length was set to 20 ms. We fine-tuned the model in a few-shot manner, involving longer fine-tuning, more evaluation steps during training, and early stopping with 20 epochs based on validation loss. To ensure a fair comparison with previous studies, we employ either off-the-shelf software packages or utilize the code provided by respective authors. Each model underwent ten executions, and the average performance across these runs is considered the final result. The hyper-parameters are chosen as default if provided, or tuned otherwise. We evaluated the models using weighted accuracy (WA) and unweighted accuracy (UA) (Metallinou et al., 2010) in speaker-independent settings. We did not require the data to be labeled by actual annotators. Instead, we used the ground-truth labels available in the training dataset. Specifically, we masked the labels and only received them when the active learning methods determined that the samples should be labeled. This approach is a common technique used by active learning researchers to validate their methods (R.Nicholas and M.Andrew, 2001). However, it is worth mentioning that, in real-world scenarios, human annotators are responsible for labeling the data.

Initialization for Active Learning. We observe that active learning methods are particularly sensitive to initialization, as the initial set of samples can substantially impact the selection order of subsequent samples in each iteration of AL. However, most active learning methods randomly select 1% samples for initialization (Margatina et al., 2021). In contrast, we propose a novel clustering-based (K-means) initialization method to improve SER performance. Specifically, we first extract sample representations of the training data from the wav2vec 2.0 CNN-based encoder. Then, we employ K-means in the training data and select 1% of the samples closest to the cluster centers as our initialized samples. It is important to note that we use the elbow method (Sammouda and El-Zaart, 2021) to automatically determine the

[3] https://huggingface.co/facebook/wav2vec2-base

number of clusters for K-means, and we use the Euclidean distance to measure the distance between sample representations.

3.4.4 Active Learning Strategies Selection for AFTER

As shown in Figure 3.1 (c), AFTER incorporates an active learning strategy for sample selection. To identify the most suitable active learning method for AFTER, we combined it with multiple well-known active learning methods and evaluated their performance. Furthermore, we find that active learning methods are sensitive to initialization, with most active learning methods randomly selecting 1% samples for initialization (Margatina et al., 2021). Unlike them, we proposed a novel clustering-based (K-means) initialization method to improve the performance of SER. Specifically, we first extract sample representations of training data from the wav2vec 2.0 CNN-based encoder. Then, we employed K-means on the training data and selected 1% of samples closest to the cluster centers as our initialized samples. Please note that we use the elbow method (Sammouda and El-Zaart, 2021) to determine the number of clusters for K-means automatically, and we use the Euclidean distance to measure the distance between sample representations.

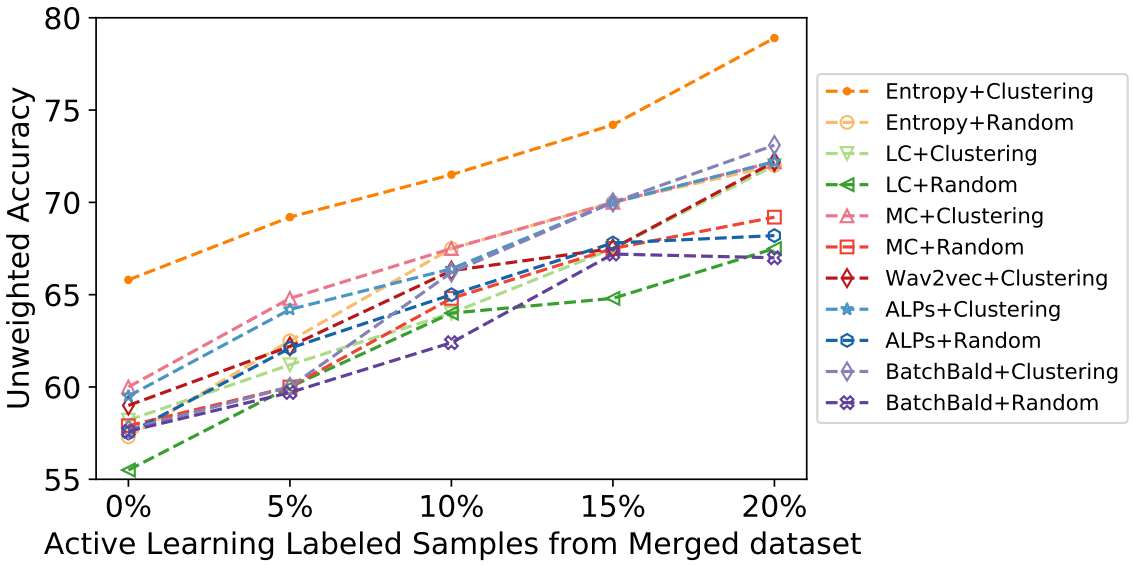


Figure 3.2: Ratio of labeled samples vs. Unweighted Accuracy.

Figure 3.2 shows that clustering-based initialization outperformed random initialization for all active learning methods. The initial set of samples significantly influenced the selection order of the samples in each iteration of AL, and effective initialization improved the performance and stability of active learning methods. Figure 3.2 illustrates that *Entropy+Clustering* emerged as the most effective active learning strategy for AFTER on the Merged-I dataset. Although we only displayed the diagram for UA due to space constraints, the diagram for WA exhibited similar trends. **Therefore, *Entropy+Clustering* was selected as the primary active learning method for AFTER**. We recommend using *Entropy+Clustering*, the simplest yet most efficient strategy for real-world applications.

Table 3.2: After with Entropy to select 10%-100% labeled samples of the Merged-I dataset for fine-tuning.

Datasets	***AFTER (TAPT+ AL-based FT)***					
	10%	20%	40%	60%	80%	100%
UA	71.45	**77.41**	78.64	**79.32**	79.26	79.15
WA	69.01	**74.32**	75.48	**76.03**	75.92	75.94
Time (mins)	262.8	**316.4**	785.4	**942.2**	1182.6	1508.2

We analyzed the relationship between the ratio of labeled samples, performance, and time consumption of AFTER. Results in Table 3.2 show that both performance and time consumption of AFTER increased as the ratio of labeled samples increased. **Our findings indicate that using 20% labeled samples yielded a significant improvement in performance while reducing the time consumption by 79% compared to fine-tuning on 100% samples.** Thus, we selected 20% labeled samples as a trade-off between performance and time consumption for subsequent experiments.

3.5 Experimental Results and Discussion

3.5.1 Comparison with Other Initialized Strategies

As illustrated in Section 3.4.3, we propose a simple but efficient method of initializing K-means that is applicable to various SER tasks. Although representative methods such

as BMAL (Chakraborty et al., 2015) and density-based methods such as DACS (Kim and Shin, 2022) are also available, our focus remains on demonstrating the effectiveness of our proposed initialization method. To assess the effectiveness of our approach, we compare it with BMAL, DACS, and random sampling as baseline initialization methods. Specifically, the selected initialization baselines are introduced as follows:

(1) DACS (Kim and Shin, 2022) is a density-aware Core-set approach used to estimate sample densities and selectively choose diverse points from sparse regions. For each input $\boldsymbol{x}_i$, the density score for each sample is calculated as follows:

$$\text{Density}(\boldsymbol{x}_i) = \frac{1}{k} \sum_{j \in \mathcal{N}(\boldsymbol{x}_i, k)} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2, \tag{3.10}$$

where $\mathcal{N}(\boldsymbol{x}_i, k)$ represents the k-nearest neighbors of $\boldsymbol{x}_i$ (Kim and Shin, 2022). We use default parameters from their original paper, selecting the top 1% of samples with the highest scores as initialized samples for downstream tasks.

(2) BMAL (Chakraborty et al., 2015) considers the distance between a sample and its surrounding labeled samples to enrich the diversity of the labeled dataset. Diversity is measured by the KL-divergence of the class probabilities distribution of similar neighboring instances, formulated as:

$$\text{Divergence}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_j P(\hat{y}_j|\boldsymbol{x}_i) - P(\hat{y}_j|\boldsymbol{x}_j) \log \frac{P(\hat{y}_j|\boldsymbol{x}_i)}{P(\hat{y}_j|\boldsymbol{x}_j)}, \tag{3.11}$$

where $\hat{y}_j$ is the predicted label for the j-th sample. We use the default parameters from their original paper, selecting the top 1% of samples with the highest scores as initialized samples for downstream tasks.

As depicted in Figure 3.3, we observe that K-means exhibits comparable performance to DACS and outperforms BMAL and Random Sampling. DACS operates as a density-sensitive core-set approach. When selecting 1% diverse samples for initialization, both K-means and DACS tend to choose the same sample set nearest to each clustering center. In contrast, BMAL and Random Sampling struggle to select the most representative samples, leading to decreased classification performance. Instead of employing DACS with its numerous

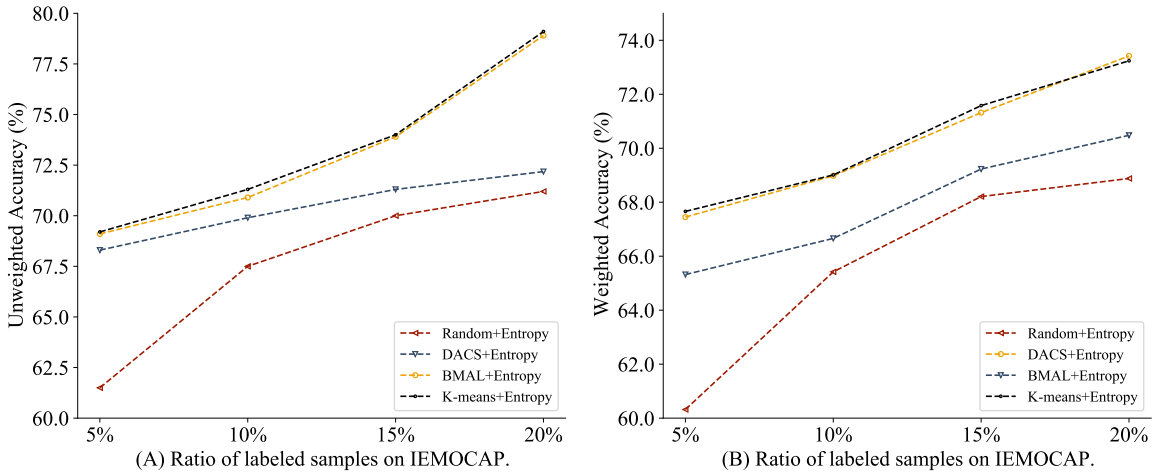


Figure 3.3: Comparison of various initialization methods for AL, with Entropy employed as the active learning strategy. Initialization involves selecting 1% of the samples.

hyperparameters requiring adjustment, we opted for the simple yet efficient K-means as our initialization method. Furthermore, our future work will explore the utilization of more representative-based methods for initialization.

3.5.2 Comparison with Best-performing Baselines

Table 3.3 displays the main results of AFTER and the baseline methods on three datasets (IEMOCAP, Merged-I dataset, and Merged-II dataset) in terms of UA and WA. To provide a more detailed comparison of baseline performance, we outline the backbone architecture, backbone size, and pre-training datasets of baselines in Table 3.4. AFTER demonstrated superior performance compared to all baselines, achieving this with only 20% labeled samples for fine-tuning, whereas existing baselines use entire datasets for training. Specifically, on the IEMOCAP dataset, AFTER improved UA and WA by **2.38%** and **0.36%**, respectively, compared to the SOTA baseline (UA of Pesudo-TAPT and WA of GLAM). Furthermore, in the Merged-I dataset, AFTER improved UA and WA by **8.45%** and **4.12%**, respectively, compared to the SOTA baseline (UA of GLAM and WA of Light). Regarding the Merged-II dataset, AFTER showed improvements of **8.30%** and **5.84%** in UA and WA, respectively,

[4]https://www.openslr.org/12
[5]https://github.com/facebookresearch/libri-light

Table 3.3: Overall performance comparison on 4 emotion categories. AFTER adopted Entropy+Clustering and selected 20% samples for fine-tuning. Baselines use all samples from each corresponding training data for training. The symbol † indicates that AFTER significantly surpassed all baselines with $p < 0.05$ according to the t-test.

Methods	**IEMOCAP**		**Merged dataset**		**Merged-2 dataset**	
	UA ↑	WA ↑	UA ↑	WA ↑	UA ↑	WA ↑
GLAM [2022]	74.01	72.98	71.38	69.28	70.21	68.32
LSSED [2021]	73.09	68.35	25.00	36.20	22.35	33.47
RH-emo [2022]	68.26	67.35	43.20	42.80	42.18	40.78
Light [2022]	70.76	70.23	69.28	71.38	68.36	70.27
Pseudo-TAPT [2022]	74.30	70.26	71.25	68.83	70.38	68.74
w2v2-L-robust [2023]	74.28	70.23	71.22	68.77	71.64	68.98
AFTER	**76.07**†	**73.24**†	**77.41**†	**74.32**†	**77.59**†	**74.38**†

compared to the UA of w2vw-L-r-12 and WA of Light.

Based on Tables 3.3 and 3.4, we have four findings to share as follows:

(1) **Larger-scale pre-training models yield better performance:** Traditional CNN-based backbones were insensitive to the pre-training process, as evidenced by GLAM (without pre-training) outperforming LSSED and RH-emo (pre-trained with ResNet). Conversely, larger-scale wav2vec 2.0-based pre-training methods, such as pseudo-TAPT and w2v2-L-robust, significantly outperformed CNN-based models, benefiting from a broader range of hyperparameters and larger pre-training datasets. Even when employing the same backbone as pseudo-TAPT, our method AFTER outperformed pseudo-TAPT and w2v2-L-robust on all three datasets, demonstrating the effectiveness and applicability of active learning for SER.

(2) **Larger-scale pre-training models exhibit denoising capabilities to a certain extent:** wav2vec 2.0-based methods significantly outperformed CNN-based methods on the Merged-I dataset and Merged-II dataset, while showing comparable performance on the IEMOCAP dataset. LSSED (Fan et al., 2021) and RH-emo (Guizzo et al., 2022) achieved favorable results with IEMOCAP but showed poor performance with the Merged-I dataset and Merged-II dataset, possibly due to their limited denoising and

Table 3.4: Comparison of baseline architectures. All information is reported by summarizing from their original papers.

Methods	Backbone	Backbone Size	Pre-training datasets
GLAM [2022]	CNNs	15 M	Without pre-training
LSSED [2021]	ResNet152	60 M	ImageNet-1k
RH-emo [2022]	ResNet50	25 M	ImageNet-1k and IEMOCAP
Light [2022]	CNNs	7 M	Without pre-training
Pseudo-TAPT [2022]	wav2vec 2.0 Base	94 M	Librispeech [4] and IEMOCAP
w2v2-L-robust [2023]	wav2vec 2.0 Large	317 M	Librispeech and Libri-Light [5]
AFTER	wav2vec 2.0 Base	94 M	Librispeech and IEMOCAP

domain transfer capabilities. In contrast, GLAM (Zhu and Li, 2022) and Light (Aftab et al., 2022) employ multi-scale feature representations and deep convolution blocks to capture high-level global data features, advantageous for filtering out noisy low-level features and enhancing performance across all datasets. Pseudo-TAPT, w2v2-L-robust, and AFTER adopt larger-scale pre-training models, which understand relevant features for the downstream SER task and help denoise irrelevant or noisy features to improve robustness against real-world datasets.

(3) **Active Learning can help achieve better performance on real-world datasets:** AFTER achieved superior classification accuracy when utilizing the Merged-I dataset and Merged-II dataset compared to solely relying on the IEMOCAP. However, baselines achieved their optimal performance solely with the IEMOCAP, as they are susceptible to the influence of outliers and redundant data. Pseudo-TAPT (L.Chen and A.Rudnicky, 2022) enhances model robustness by using K-means to capture higher-level frame emotion labels as pseudo labels for supervised TAPT. Although baselines can mitigate dataset noise to a certain extent, they exhibit high time complexity during fine-tuning with large-scale datasets and fail to effectively bridge the gap between pre-training and the downstream SER task. In contrast, AFTER uses unsupervised TAPT to mitigate the information gap between the source domain (ASR) and the target (SER) domain. Additionally, AFTER selects a subset of the most informative and diverse samples for iterative fine-tuning, offering three advantages: Firstly, it

reduces labor consumption for manually labeling large-scale SER samples; Secondly, by utilizing a smaller labeled dataset, AFTER significantly reduces the overall time consumption (Figure 3.5), making it practical and feasible for real-world applications; Finally, the iterative fine-tuning process employed by AFTER improves performance and stability by eliminating noise and outliers present in the selected samples, leading to enhanced overall model performance in SER tasks.

(4) **Baselines performed better on the Merged-II dataset than on the Merged-1 dataset:** AFTER achieved superior classification performance on the Merged-II dataset compared to the Merged-I dataset. However, the combination of acted and spontaneous speech datasets posed greater challenges for other baselines due to their sensitivity to heterogeneous and noisy samples. Merging multiple datasets enabled AFTER to extract a wider variety of samples from a larger pool of data. This increased diversity of samples contains more information, improving classification performance.

Table 3.5: Overall performance comparison on the SAVEE dataset with seven emotion categories. AFTER adopted Entropy+Clustering and selected 20% samples for fine-tuning. We obtained the baselines' performance directly from TIM-Net.

Methods	UA ↑	WA ↑
DCNN [2020]	-	82.10
TSP+INCA [2021]	83.38	84.79
CPAC [2022]	83.69	85.63
GM-TCN [2022]	83.88	86.02
TIM-Net [2023]	86.07	87.71
AFTER	**86.23**	**87.98**

As depicted in Table 3.5, AFTER demonstrated superior classification performance on the SAVEE dataset with seven emotion categories, underscoring its capacity to recognize a wider spectrum of emotions. Specifically, AFTER improved UA and WA by **0.19%** and **0.31%**, respectively, using only 20% of the samples. By iteratively extracting the most informative and uncertain samples, AFTER fine-tunes the SSL model wav2vec 2.0, effectively removing irrelevant samples and outliers, thereby improving classification performance.

Table 3.6: Weighted Accuracy comparison on seven emotion categories. AFTER adopted Entropy+Clustering and selected 20% samples for fine-tuning. Baselines use all samples from each corresponding dataset for training.

Methods	**BAUM-1s**	**Merged-3 dataset**
MFCC+PLP+SVM (Zhalehpour et al., 2017)	29.41	28.54
CNN+SVM (Zhang et al., 2018b)	42.28	40.68
CNN+DTPM+SVM (Zhang et al., 2018a)	44.67	42.33
CNN+SVM (Ma et al., 2019)	42.39	41.79
CNN+LSTM (Zhang et al., 2022a)	50.22	48.39
AFTER	**50.64**	**51.24**

As shown in Table 3.6, we also assess the performance of AFTER on BAUM-1s and Merged-III dataset (containing spontaneous datasets) with seven emotion categories (Evaluation on BAUM-1s (Zhang et al., 2022a)). We observed that AFTER can achieve SOTA performance on both BAUM-1s and Merged-III dataset. AFTER obtained better performance on the Merged-III dataset by selecting more diverse samples from a large pool of datasets. Baselines lacked the ability to remove noisy data from the merged dataset, decreasing their performance on real-world scenarios.

3.5.3 Ablation Study for AFTER

We performed an additional ablation study to assess the efficacy of AFTER, as shown in Table 3.7. Specifically, we conducted fine-tuning (FT) and TAPT+FT on random sample selection and AL-based (Entropy) sample selection with varying ratios of labeled samples, ranging from 10% to 100%. To ensure a fair comparison, we adopted the same K-means initialization for them and other hyperparameters, *i.e.*, learning rate and random seeds.

From Table 3.7, we have four interesting observations:

(1) Fine-tuning with active learning improved performance compared to random sampling (FT+Entropy vs. FT+Random), regardless of the number of labeled samples. This result demonstrates AL-based fine-tuning strategy efficiently eliminates outliers and selects the most informative and diverse samples for fine-tuning;

Table 3.7: Ablation study on the Merged-I dataset. FT means fine-tuning, and TAPT+FT indicates the adopting of TAPT followed by fine-tuning with the corresponding selected labeled samples. AFTER adopted Entropy to select samples for fine-tuning. Random Sampling and Entropy Sampling utilize the same K-means initialization.

Methods	**Random Sampling**				**Entropy Sampling**			
	FT		**TAPT+FT**		**FT**		**AFTER**	
	UA ↑	WA ↑	UA ↑	WA ↑	UA ↑	WA ↑	UA ↑	WA ↑
10%	50.82	48.96	70.21	68.85	68.21	66.32	71.45	69.01
20%	51.37	49.92	73.82	71.33	71.07	68.21	**77.41**	**74.32**
30%	52.37	50.18	74.49	71.89	72.35	69.21	78.20	75.16
40%	55.68	52.21	76.01	72.28	73.55	70.18	78.64	75.48
60%	60.39	59.32	77.21	74.58	74.28	71.35	**79.32**	**76.03**
80%	58.34	56.72	78.88	75.82	73.52	70.39	79.26	75.92
100%	57.21	54.12	78.21	75.36	73.89	70.89	79.15	75.94

(2) TAPT+FT outperformed FT on both random sampling and Entropy sampling, indicating that TAPT can effectively minimize the domain difference and significantly enhance the performance of the downstream SER task;

(3) With the same number of labeled samples, AFTER outperforms TAPT+FT+Random on the Merged-I dataset. However, AFTER with 20% labeled samples performs worse than TAPT+FT+Random with 80%$\sim$100% labeled samples. The reason is that TAPT+FT uses more labeled data for fine-tuning to prevent the model from overfitting and improve its robustness. In a fair comparison with the same size of the training data for fine-tuning, TAPT+FT+Random with 20% labeled samples performed worse than AFTER(20%), demonstrating the effectiveness of AFTER;

(4) When 100% of samples are used, AL-based methods significantly outperforms the random sampling method (FT+Random vs FT+Entropy). The main reason is that Random sampling is affected by noise data, and the model constantly corrects the classification boundary, making it difficult to improve the results. Entropy sampling avoids the effect of noisy data by selecting the most informative and diverse samples for FT in advance to fix the classification boundary properly.

3.5.4 Visualization of AFTER

As depicted in Figure 3.4, we present qualitative comparisons of AFTER with random sampling. Our observations indicate that AFTER tends to select samples that are representative and uncertain. Specifically, AFTER selects samples near each clustering center, benefiting from the K-means initialization, which are the most representative samples of the entire datasets. Using entropy as a criterion, AFTER selects the most uncertain samples for labeling, which almost lie on the clustering boundaries. Based on previous experimental results, we discovered that only these selected representative and highly uncertain samples could achieve comparable or even superior performance compared to training with the entire datasets. Additionally, the samples selected for training via random sampling are shown in Figure 3.4 (A) and (C). We found that random sampling tends to select outliers and redundant samples (some points overlap due to repeated selection).

3.5.5 Time Consumption Comparison

Figure 3.5 (A) demonstrates that FT+AL with 20% labeled samples significantly reduced the time consumption of FT (fine-tuning on all labeled samples). Compared to TAPT+FT, TAPT+FT+AL significantly decreased time consumption with the main cost incurred by TAPT. Furthermore, the relationship between time consumption and the ratio of labeled samples is shown in Figure 3.5 (B). AL-based fine-tuning exhibited a linear increase in time consumption with sample size from 1%$\sim$20% (exponential growth from 30%$\sim$100% in Table 3.2), indicating the efficiency of AFTER and its potential to be applied in large-scale unlabeled real-world scenarios.

Figure 3.6 illustrates the relationship between running time and unweighted accuracy of classification on both the IEMOCAP and Merged-I dataset. We observe that FT+AL outperformed FT on both datasets, demonstrating the effectiveness of active learning. Additionally, we also observe that TAPT+FT+AL underperformed FT and FT+AL within the time intervals of 0 to 250 minutes on IEMOCAP and 0 to 300 minutes on the Merged-I dataset, respectively. This is because TAPT requires time to adaptively pre-train the model using downstream unlabeled training datasets. Following the TAPT process, TAPT+FT+AL

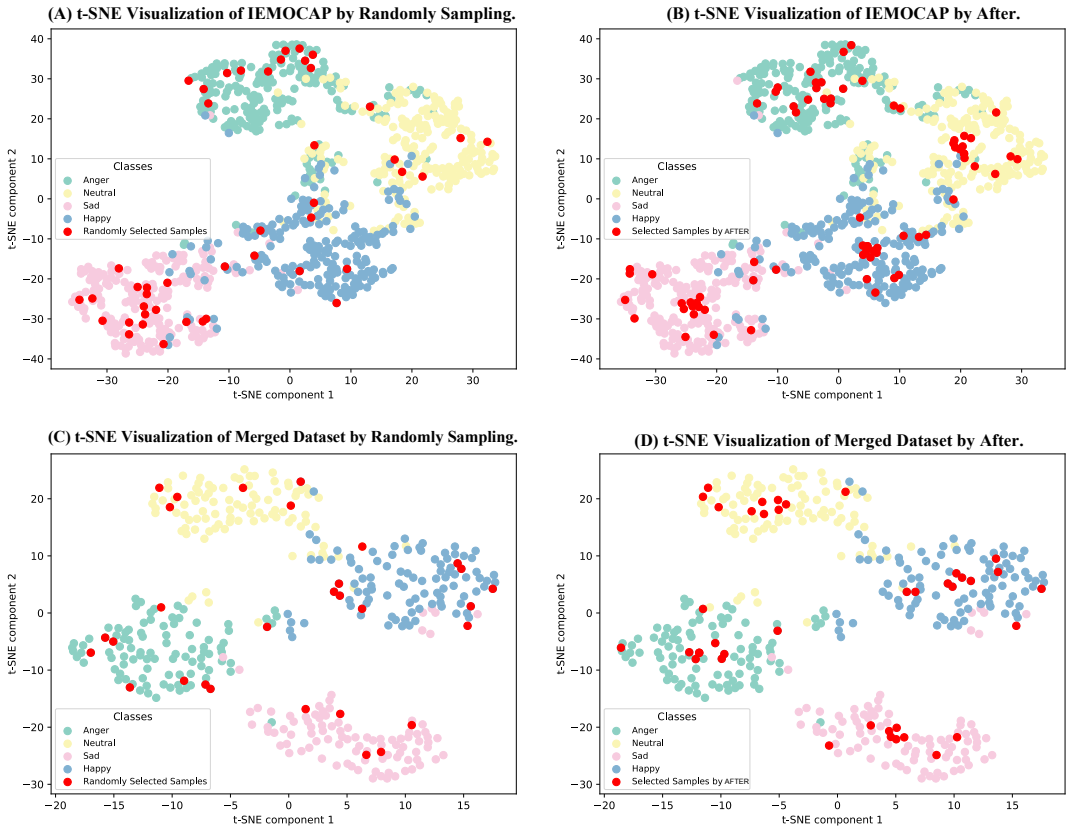


Figure 3.4: t-SNE visualization of AFTER and randomly sampled methods. The selected samples are represented with red colors on the IEMOCAP and Merged-I dataset by either randomly sampling or AFTER.

(AFTER) significantly outperformed the other two baselines on both datasets, thus demonstrating the effectiveness of our proposed method.

3.5.6 Adapting AFTER with Different Pre-trained ASR Models

Many SSL models have recently been proposed [6], such as LABERT (Fatehi and Kucukyilmaz, 2023) and DPHuBERT (Peng et al., 2023). To demonstrate the versatility of our proposed method, AFTER, in various SSL models, we replaced wav2vec 2.0 with another widely used SSL model, HuBERT (Hsu et al., 2021; Xin et al., 2023), in conducting our

[6] https://www.isca-archive.org/interspeech_2023/index.html

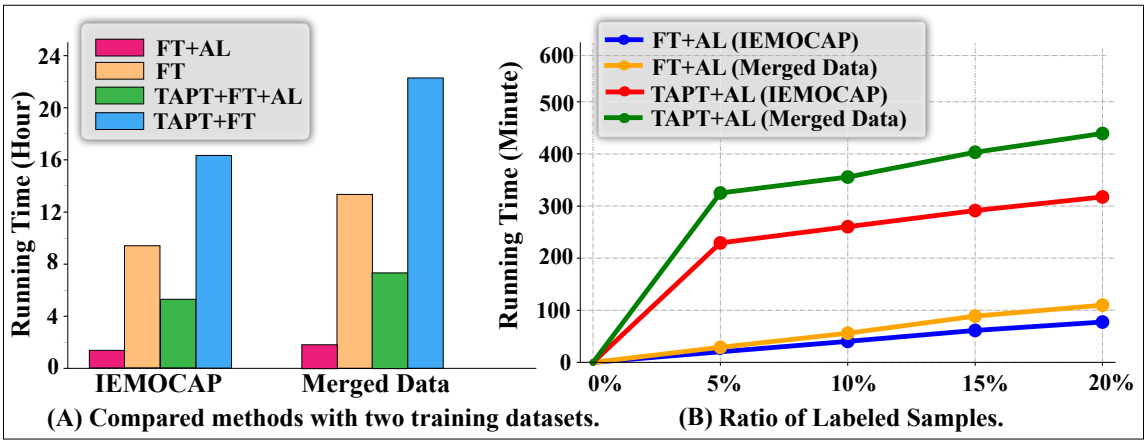


Figure 3.5: (A) Time Consumption Comparison and (B) Relationship between ratio of labeled samples and time consumption.

experiments. We first briefly introduce the HuBERT. HuBERT, also known as Hidden-Unit BERT, is a variant of BERT designed specifically for speech processing tasks. Its core techniques are summarized as follows:

- **Architecture Adaptation**: HuBERT adapts the BERT architecture to process speech data. It modifies the input layer to accommodate raw audio waveforms and adjusts subsequent layers to handle sequential data inherent in speech signals. HuBERT incorporates transformer layers to capture contextual information from the extracted features. These layers enable the model to understand the temporal dependencies and nuances present in speech sequences.

- **Feature Extraction**: Similar to traditional speech processing pipelines, HuBERT first extracts high-level features from raw audio using techniques like Mel-frequency cepstral coefficients or filter banks. These features capture important acoustic characteristics of the speech signal.

- **Training Strategy**: HuBERT is pre-trained on large-scale unlabeled speech datasets using self-supervised learning techniques. During pre-training, the model learns to predict masked portions of the input sequence or to reconstruct corrupted segments, leveraging the bidirectional context provided by the transformer architecture.

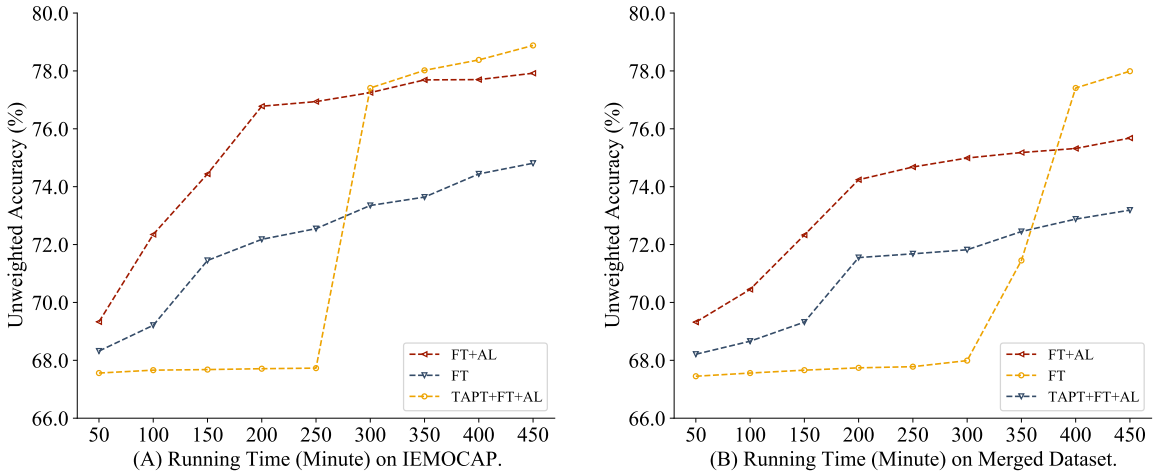


Figure 3.6: A plot illustrating the efficiency of AFTER. The x-axis represents the running time, while the y-axis indicates the unweighted classification accuracy.

The TAPT process for HuBERT-based AFTER follows a similar way to that of word2vec 2.0-based AFTER. We first pre-train HuBERT using downstream speech emotion training datasets without using emotion labels, in an SSL manner. We randomly mask 20% tokens and reconstruct them using emotion recognition training datasets. Then, similar to theword2vec 2.0-based AFTER method, we use K-means to select 1% samples for initialization. Finally, we iteratively query samples to train the HuBERT-based AFTER model.

Experimental results are presented in Table 3.8. We have two observations: (1) HuBERT-based fine-tuning with random sampling surpassed wav2vec 2.0-based fine-tuning with random sampling on the Merged-I datasets by nearly 2%, showcasing HuBERT's effectiveness. This improvement can be attributed to HuBERT's tailored architecture for processing speech data, which may better capture subtle emotional cues. Additionally, HuBERT's pre-training objectives are more aligned with the demands of emotion recognition tasks. (2) wav2vec 2.0-based AFTER outperformed HuBERT-based AFTER. This is because wav2vec 2.0 pre-trains the model with contrastive loss and reconstructing loss, which better assists the SSL model in understanding the context of downstream datasets. These observations highlight the advantages and differences between HuBERT-based and wav2vec 2.0-based fine-tuning, as well as the impact of pre-training strategies on downstream performance.

Table 3.8: Unweighted and weighted accuracy on the Merged-I dataset. Entropy is used as an AL strategy for both HuBERT and wav2vec 2.0 backbones.

Methods	**HuBERT as backbone**				**wav2vec 2.0 as backbone**			
	FT+Random		**AFTER**		**FT+Random**		**AFTER**	
	UA ↑	WA ↑	UA ↑	WA ↑	UA ↑	WA ↑	UA ↑	WA ↑
10%	52.28	50.58	71.22	68.83	50.82	48.96	71.45	69.01
20%	54.32	52.43	76.32	74.11	51.37	49.92	77.41	74.32
30%	58.44	56.26	77.54	75.12	52.37	50.18	78.20	75.16
40%	60.55	59.42	78.23	75.18	55.68	52.21	78.64	75.48
60%	63.38	61.43	78.98	75.95	60.39	59.32	**79.32**	**76.03**
80%	64.56	62.34	79.24	75.81	58.34	56.72	79.26	75.92
100%	63.45	61.28	78.89	79.35	57.21	54.12	79.15	75.94

3.6 Limitations and Summary

In this section, we give the limitations of this study and future work in Sec 3.6.1. Then, we give an overall summary of the proposed model in Sec 3.6.2.

3.6.1 Limitations

Although AFTER achieves SOTA performance in IEMOCAP and all merged dataset, there are still some limitations in this study. (1) Its performance on larger-scale and more heterogeneous real-world data remains unclear. (2) Another limitation of AFTER is the time-consuming process of calculating the entropy of each sample in each AL iteration. Additionally, we have only explored five of the most commonly used AL strategies in this study, leaving better strategies unexplored. (3) The annotation of emotions varies greatly depending on various factors. Although this study assumes that annotators can always provide ground-truth labels, current experimental settings may not sufficiently simulate real-world human-in-the-loop situations involving multiple annotators. (4) We need to apply our methods to more complicated scenes, such as social networks with the clustering technique (Li et al., 2021a, 2023a, 2021b) and clinical patient emotion detection (Li and Ma, 2019; Li et al., 2022b).

3.6.2 Summary

In this study, we investigated unsupervised TAPT and the AL-based fine-tuning strategy to improve the performance of speech emotion recognition. To extend speech emotion recognition applications to real-world scenarios, we created three large-scale noisy and heterogeneous datasets, and we used TAPT to bridge the information gap between pre-trained and the target speech emotion recognition task. Extensive experimental findings demonstrate that AFTER significantly improved performance and reduced time consumption. In our future work, we plan to create larger-scale speech emotion recognition datasets for testing in the speech domain. Furthermore, our objective is to explore and design more effective and efficient active learning strategies tailored to the speech emotion recognition task, aiming to minimize time consumption. Finally, we would like to propose a more general framework that extends beyond SER, focusing on a wider range of speech- or language-related tasks.

Part II

Multimodal Emotion Recognition

Chapter 4

Overview of Multimodal Emotion Recognition

In PART II, we describe multimodal emotion recognition: its formulation and development over recent years, the key components of multimodal emotion recognition systems, and future research directions.

Firstly, we introduce the main multimodal fusion mechanisms in Section 4.1. Specifically, we introduce the early fusion methods in Section 4.1.1, late fusion mechanism in Section 4.1.2 and hybrid fusion mechanism in Section 4.1.3.

Then, we introduce the multimodal emotion recognition backbones in Section 4.2. We introduce deep neural networks-based methods in Section 4.2.1. We introduce Seq2seq-based methods in Section 4.2.2. Transformer-based backbones are introduced in Section 4.2.3. The graph neural network-based backbones are introduced in Section 4.2.4.

Next, we introduce the main multimodal emotional datasets in Section 4.3. Finally, we introduce the main evaluation metrics in Section 4.4.

4.1 Multimodal Fusion Mechanisms

4.1.1 Early Fusion Mechanism

In the field of Multimodal Emotion Recognition (MER), early fusion, also known as feature-level fusion, is simple and has low computational complexity. This method combines features from different sources such as text, speech, and vision by joining them together. These combined features are then used as inputs for training deep neural networks (DNNs). For example, Poria et al. (2015) were pioneers in implementing this fusion method. They extracted feature vectors from speech, visual, and text modalities, merged them into a comprehensive feature vector, and then used this unified vector as input for the classification model. This work laid the foundation for further exploration of early fusion. Inspired by their work, Huang et al. (2018) improved the data richness to test the effectiveness of early fusion. During the 2018 Audiovisual Emotional Challenge, they increased the number of training samples by creating shorter samples that overlap the original ones. This increased data diversity, combined with early fusion, showed the potential of this approach. Williams et al. (2018) made further progress in early fusion. They created a system that can recognize emotions and their intensity by combining features from different sources at the input level. They used a bidirectional long short-term memory model for sequence learning. This showed that early fusion can effectively predict the presence and rough intensity of multiple emotions. The main advantage of early fusion is its ability to quickly combine different features, which significantly improves the performance of the model in recognizing emotions. This makes early fusion valuable in MER applications. However, early fusion also has limitations. Simply joining data from different sources might prevent the model from properly handling the unique aspects of each source, leading to missing some important emotional information. Additionally, this method cannot effectively filter out conflicting or redundant information from different sources. Early fusion also struggles with synchronizing data from different sources, as they often have different time scales and structures, making it difficult to align them. In summary, while early fusion is important in MER, we need to address its limitations to improve the method. This ongoing research is crucial for the continuous development of MER.

4.1.2 Late Fusion Mechanism

Late fusion, also known as decision-level fusion, is commonly used in multimodal emotion recognition. The basic idea is to independently extract features and train models for each modality and then combine their prediction results using strategies such as averaging, weighted sum, majority voting, or deep neural networks. Due to its feature-independent fusion process and uncorrelated mistakes from different classifiers, this fusion method has garnered significant attention. Several researchers have demonstrated the effectiveness of late fusion. For example, Poria et al. (2016) in 2016 developed separate models for speech, text, and visual modalities. After training, these models generated prediction probabilities for each modality, which were then combined using weighted fusion to produce the final prediction. Similarly, Huang et al. (2017b) used Long- and Short-Term Memory networks to create models that capture emotional information from each modality. They used Support Vector Regression to combine the predictions at the decision level. Su et al. (2020) proposed a multi-level segmented decision-level fusion model for emotion recognition. They used bidirectional LSTM to learn emotional features and SVR to combine the predictions. Bidirectional LSTM modeled different emotional information, considering the influence of past and future features, while SVR helped reduce redundant information. Recently, Sun et al. (2020) used LSTM and a self-attention mechanism to capture complex temporal dependencies in feature sequences. They then used a second-level LSTM to combine the predictions of several unimodal emotion recognition models. Compared to feature-level fusion, decision-level fusion is simpler and more flexible. It does not require the temporal synchronization of the modality features. Additionally, during the late fusion process, each modality can utilize the most suitable classifier or model to learn its features, thus improving local decision outcomes. However, late fusion assumes that each modality works independently, thereby ignoring interactions between different modalities, which can limit the accuracy of the final emotional prediction. Future research should explore methods that combine the strengths of late fusion with inter-modality interactions to improve affect recognition accuracy.

4.1.3 Hybrid Fusion Mechanism

Unlike the previously discussed early and late fusion methods, hybrid multimodal fusion combines the strengths of early and late fusion. The goal of this fusion method is to maximize the use of emotional information extracted from multimodal data while also considering the interplay and synergy between different modalities. For example, Wöllmer et al. (2013) proposed a hybrid fusion method using bidirectional long and short-term memory to merge audio and visual features at the feature level. They then combined these fused features with predictions from a text classifier using decision-level fusion, a strategy typical of late fusion methods. This approach benefits from the high-level feature interactions of early fusion and the specialized classifiers of late fusion. In Music Emotion Recognition, Nemati et al. (2019) introduced a similar hybrid fusion method. First, they used feature-level fusion to map speech and visual data to a shared latent space, similar to the early fusion. These latent features were used for emotion classification. At the same time, they trained a separate model using text features. Finally, they applied decision-level fusion, based on the Dempster-Shafer theory, to combine the text and audio-visual classification results. This method uses inter-modality interactions early on while keeping the flexibility of independent classification later. In summary, hybrid multimodal fusion aims to combine the benefits of early and late fusion methods to create a more robust and comprehensive emotion recognition system. However, these methods are complex and challenging, especially in finding the right mix of early and late fusion strategies and managing increased computational demands. Therefore, ongoing research in hybrid fusion methods is crucial to advance multimodal emotion recognition.

4.2 Multimodal Emotion Recognition Backbones

4.2.1 Deep Neural Networks-based Models

Deep neural networks backbones always independently analyze the input features of different modalities and directly concatenate these features as the input of the next layer network to learn the interaction information between modalities. For example, Nguyen et al. (2019) proposed a novel multimodal sentiment recognition model that utilizes three convolution neural networks to separately process the low-level features of the speech, text, and visual

modalities. The aim is to achieve high-level emotional feature representations. Subsequently, the feature vectors from each modality are merged into a single vector and injected into deep neural networks for emotion classification. Similarly, Ortega et al. (2019) proposed a new deep neural network for multimodal fusion emotion recognition in the audio, video, and text modalities. This deep neural network architecture includes independent and shared layers, with the aim of learning the representations of each modality as well as the optimal combined representation to achieve the best prediction performance. The experimental results achieved using the AVEC wild emotion analysis dataset indicate that the proposed deep neural network model outperforms the early and late fusion approaches. Yu et al. (2021) proposed a deep neural network-based multitask multimodal sentiment analysis network (Self-MM). In the Self-MM model, representations from different modalities are fused using the simplest concatenation method. Despite the apparent similarity between simple concatenation fusion and early fusion, there are notable distinctions between the two. Furthermore, Han et al. (2021b) pointed out that the aforementioned methods lack control in the information flow from raw input to fused embedding, which may lead to information loss and an increased risk of unintended noise. To address this issue, they introduced the concept from information theory, mutual information (MI), and proposed a hierarchical mutual information maximization framework for MER. By enhancing the MI between multimodal inputs, it can effectively filter out modality-specific noise irrelevant to the task while preserving modality-invariant content across all modalities.

4.2.2 Sequence to Sequence-based Models

Seq2Seq models adopt recurrent neural networks to deal with merged multimodal vectors for multimodal emotion recognition. For example, Chen et al. (2017) first based the idea of word-level modality fusion to propose a gated multimodal embedding LSTM with temporal attention model, which consists of two modules, *i.e.*, gated multimodal embedding and LSTM with temporal attention. The gated multimodal embedding alleviated the difficulty of fusion when noisy modalities were present. The LSTMs with temporal attention performed word-level fusion between input modalities at a finer interaction and focused on

the most significant time steps. Zadeh et al. (2018c) proposed a novel network for understanding human communication based on word-level features called Multi-Attention Recurrent Network (MARN). MARN includes two key components: long-short-term mixed (LSTHM) memory and multi-attention blocks (MAB). Hybrid memory LSTHM was an extension of LSTM that carries mixed information by reconstructing memory components. MAB was used to discover cross-view dynamics across different modalities. LSTHM was expanded at the word level in time steps, while MAB was applied at each time step to achieve fine-grained interactions between the modalities. Subsequently, Zadeh et al. (2018a) proposed a memory fusion network (MFN) for multi-view sequential learning, which also used word-level features as interactive features. The MFN first used the LSTMs module system to model the intra-modality interaction, then used the delta-memory attention network (DMAN) module to achieve fine-grained inter-modality interactions. Conversational memory network (Hazarika et al., 2018) leverages contextual information from the conversation history and uses gated recurrent units to model past utterances of each speaker into memories. DialogueRNN (Majumder et al., 2019) that proposes an attention mechanism over the different utterances and models emotional dynamics by its party GRU and global GRU. bc-LSTM (Poria et al., 2017) proposes an LSTM-based model that captures contextual information from the surrounding utterances.

4.2.3 Transformer-based Models

Self-attention mechanism, used in Transformers, has achieved great success in recent years. Thus, Transformer-based multimodal emotion recognition frameworks have been proposed. For example, Wu et al. (2022) developed a fusion network named Bimodal Information-augmented Multi-Head Attention, which relies on multi-head attention and comprises a total of four layers. The first layer captures modality-specific dynamics within a single view. The second layer represents cross-view dynamics. In the third layer, bimodal interaction information is extracted by utilizing a multi-head attention mechanism, which calculates bimodal attention for weight allocation in feature attention. Lastly, the independent modality embeddings are concatenated with bimodal attention to form a multimodal representation used for predicting the emotion of each utterance. Tsai et al. (2019a) proposed a multimodal

Transformer for unaligned multimodal language sequences. The core of the Mult model is cross-modal attention, which attends to fine-grained interactions between multimodal sequences across distinct time steps and latently adapts streams from one modality to another. Although the multimodal Transformer method extends the self-attention mechanism of the original Transformer network to learn cross-modal dependencies between elements, directly copying the self-attention is influenced by the mismatch between different modal features, leading to potentially unreliable cross-modal dependencies. Based on this observation, Liang et al. (2021) introduced the Modality-Invariant Cross-Modal Attention (MICA) method, which learns cross-modal interactions in a modality-invariant space and effectively solves the sequence matching problem in unaligned features. The effectiveness of the MICA method has been validated through experiments on multiple datasets. Additionally, to tackle the problem of cross-modal asynchrony in multimodal sequences, Lv et al. (2021) introduced the Progressive Modality Reinforcement (PMR) method, which relies on cross-modal Transformers. This method incorporates a message hub that facilitates information exchange among modalities. The message hub sends mutually shared messages to each modality, reinforcing their features through cross-modal attention. Simultaneously, the message hub accumulates reinforced features from each modality and leverages them to generate enhanced common messages. This iterative process enables the gradual complementary integration of shared information and modality-specific features. Ultimately, the enhanced features are employed for sentiment prediction tasks. Tzirakis et al. (2021) presented a textual architecture based on Transformers and an attention-based fusion strategy to effectively integrate diverse modal features and enhance sentiment recognition performance. The proposed textual model employs multilinear projection and context-aware feature generators to capture sentence semantics. Moreover, the proposed fusion strategy achieves superior balance among the relationships across different modalities compared to a straightforward concatenation approach, resulting in enhanced recognition performance. In summary, fine-grained interaction fusion methods provide an innovative avenue for capturing nuanced interactions between modalities, proving to be pivotal in advancing MER. These methods ensure that both global and local information is integrated for precise emotion recognition. While advancements have been made in this sphere, continuous research is warranted to refine these methods further and develop more robust systems for MER. More recently,

Lian et al. (2023) propose a novel framework that combines semi-supervised learning with multimodal interactions. However, it currently addresses only two modalities, *i.e.*, text and audio, with visual information reserved for future work. Shi and Huang (2023) introduces MultiEMO, an attention-based multimodal fusion framework that effectively integrates information from textual, audio and visual modalities. However, neither of these models addresses the temporal aspect in conversations.

4.2.4 Graph Neural Networks-based Methods

Recently, more and more attention has been focused on graph-based multimodal emotion recognition since graph structure is suitable for extracting inter- and intra-person dependencies and improve the performance of multimodal emotion recognition. For example, Zadeh et al. (2018b) introduced a novel multimodal fusion technique known as Dynamic Fusion Graph (DFG) to investigate the interaction between modalities in human multimodal language. DFG is an improved version of MFN in which the original fusion method has been replaced with DFG. CONSK-GCN (Fu et al., 2021) uses graph convolutional network (GCN) with knowledge graphs. Lian et al. (2020) use GNN based architecture for Emotion Recognition using text and speech modalities. Zhang et al. (2019) models utterances and speakers as nodes in a graph, capturing context dependencies and speaker dependencies as edges. However, ConGCN focuses only on textual and acoustic features and does not consider other modalities. MMGCN (Wei et al., 2019), on the other hand, is a graph convolutional network (GCN)-based model that effectively captures both long-distance contextual information and multimodal interactive information.

4.3 Multimodal Emotion Recognition Datasets

In this section, we introduce the datasets for multimodal emotion recognition as follows.

4.3.1 Popular Multimodal Datasets

(1) **CMU-MOSI.** The Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) dataset, introduced by Zadeh et al. (2016a), marks a significant milestone in multimodal sentiment analysis databases due to its unique characteristic of incorporating subjective sentiment and emotional intensity annotations. The dataset includes 93 randomly selected videos from YouTube, involving 89 different speakers 41 females and 48 males. All these videos were recorded in different settings, some of which used high-tech microphones and cameras while others used less professional recording equipment. In addition, the distance between the users and the camera, as well as the background and lighting conditions, were different. The original quality of the videos remains unaltered, preserving their fidelity. This authentic approach guarantees the data accurately reflect the varying audio-visual quality in user generated content, thereby providing a robust training set for real-world applications. The CMU-MOSI dataset covers a wide spectrum of topics, including movie and book reviews and product evaluations. The videos are segmented into 2,199 clips, each rated on an emotional polarity scale, from +3 (signifying strongly positive sentiment) to -3 (signifying strongly negative sentiment). This scale was developed based on the annotations of five separate annotators, with the average of these five scores serving as the definitive emotional polarity for each clip. This methodology effectively distills the sentiments down to two categories, positive and negative, simplifying the complexity of emotions and providing a clear benchmark for training sentiment analysis models.

(2) **NNIME.** The NTHU-NTUA Chinese Interactive Emotion Corpus (NNIME) (Chou et al., 2017) leverages dyadic spoken interactions to evoke genuine emotional responses. The database features spontaneous dyadic spoken interactions involving 44 participants. These interactions have been assiduously captured, producing roughly 11 h of continuous and synchronized data across audio, video, and electrocardiogram modalities. The multimodal character of this dataset extends the scope of sentiment analysis by merging physiological indicators with audio–visual data, setting a precedent for more detailed and holistic emotion assessments. One of the salient characteristics of the NNIME dataset is its extensive annotation process. A robust

team of 49 annotators meticulously annotated the data, yielding valuable emotional insights. The annotation procedure includes four distinctive perspectives: peer reports, director reports, self-reports, and observer reports. Structurally, the NNIME dataset comprises 6,701 sentences and includes both discrete and continuous emotion labels. Discrete labels encompass emotions like anger, sadness, joy, surprise, neutrality, and happiness. Conversely, continuous labels entail the scales of valence and arousal. This two-pronged labeling system captures the subtlety of emotional expression, facilitating a more precise exploration of emotional complexities and variations.

(3) **CMU-MOSEI.** The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset (Zadeh et al., 2018b) is a crucial resource in multimodal sentiment analysis and emotion recognition, distinguished by its scale, which is unmatched in the field. This dataset encompasses 3,837 videos collected from over 1,000 distinctive YouTube speakers, with a reasonably balanced gender distribution of 57% male and 43% female speakers. One of CMU-MOSEI's unique characteristics is its annotations at the utterance level, with 23,259 samples annotated in this manner. The thematic diversity of the CMU-MOSEI dataset is notable, with a spread across 250 different themes. The three predominant themes are comments (16.2%), debates (2.9%), and consultations (1.8%), while the rest of the themes are nearly evenly distributed. Such thematic richness encourages a comprehensive contextual understanding, enabling a more in-depth exploration of sentiment and emotional expressions across varied discussion scenarios.

(4) **IEMOCAP.** The Interactive Emotional Dyadic Motion Capture dataset (Busso et al., 2008) amalgamates diverse modalities such as video, speech, facial motion capture, and text data, enabling comprehensive examination of emotional states in interactive circumstances. Ten actors, an equal mix of males and females, contributed to the data collection. These actors, paired by gender and divided into five groups, performed both scripted and improvised dialogues. This assortment of dialogues enriches the dataset with a variety of emotional content. IEMOCAP comprises 4,784 improvised and 5,255 scripted conversations, providing a wide spectrum of emotional contexts for analysis. The dialogues span nine discrete emotions (happiness, sadness, anger, surprise, fear,

disgust, frustration, excitement, and neutrality), in addition to continuous emotional dimensions like activation, arousal, and dominance, facilitating a more dynamic emotional analysis.

(5) **MELD.** The Multimodal EmotionLines Dataset (Poria et al., 2019a) offers a unique perspective by focusing on the emotional intricacies present in multi-participant dialogue. The dataset is an organized collection of dialogues from the popular American television series "Friends", featuring 1,433 conversations that account for a total of 13,708 utterances. Each utterance within the MELD dataset is annotated with one of seven emotional labels: anger, disgust, sadness, joy, neutrality, surprise, or fear, providing a diverse emotional spectrum for analysis. In addition to these discrete emotion labels, each utterance also includes an emotional classification of positive, negative, or neutral, thereby facilitating a broader understanding of the sentiment behind each statement. The dataset's design aligns with its primary objective: to furnish training data that are conducive for developing and refining contextual models in dialogue-based emotion recognition.

4.3.2 Other Multimodal Datasets

(6) **CREMA-D.** CREMA-D contains 7,442 original clips from 91 actors (Cao et al., 2014). These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities. Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral and Sad) and four different emotion levels (Low, Medium, High and Unspecified). Participants rated the emotion and emotion levels based on the combined audiovisual presentation, the video alone, and the audio alone. Due to the large number of ratings needed, this effort was crowd-sourced and a total of 2443 participants each rated 90 unique clips, 30 audio, 30 visual, and 30 audio-visual. 95% of the clips have more than 7 ratings.

(7) **Youtube.** The YouTube dataset (Morency et al., 2011) comprises 47 videos, with 20 videos featuring female subjects and the remaining 27 highlighting male perspectives.

The dataset further extends its diversity with an age range spanning from 14 to 60, depicting emotive expressions across various stages of life. Although the participants originate from a variety of cultural backgrounds, each video within the dataset is characterized by participants expressing their views in English. This complexity is crucial to the development of models that can robustly function in real-world scenarios where noise is omnipresent. The emotional labeling of the videos in the dataset further augments its value. Each of the 47 videos has been tagged with either positive, negative, or neutral emotion tags. These categorical annotations provide a solid foundation for supervised learning approaches in deep learning. The YouTube dataset contains diverse scenes and environmental noise, which makes it advantageous for improving model generalization.

(8) **MOUD.** The Multimodal Opinion Utterances Dataset (Pérez-Rosas et al., 2013) offers a unique perspective by focusing on opinion-based utterances, which are intrinsically rich in emotional cues. The MOUD consists of 498 utterances, all painstakingly labeled as either positive, negative, or neutral in emotion. An intriguing aspect of MOUD is its source; the dataset comprises videos curated from YouTube, one of the world's largest platforms for user-generated content. This provides the advantage of real-world, unscripted emotional expressions, a factor that significantly boosts the applicability and robustness of models trained on this data. The dataset features speakers whose ages range between 20 and 60 years, ensuring a good representation of emotional expressions across different age groups. It includes 15 female speakers, further enhancing its gender diversity. The average duration of the speeches in the dataset is approximately 5 s, which is a typical duration for a single utterance. The dataset includes 182 positive, 231 negative, and 85 neutral utterances.

(9) **OMG.** The One-Minute Gradual-Emotion Recognition dataset (Barros et al., 2018), which incorporates videos from a multitude of YouTube channels, provides a comprehensive view of various emotional behaviors within different contexts. The unique selection process of these videos, centered around the term "monologue", is designed to depict the gradual evolution of emotions. Every video within the OMG dataset is segmented based on speech, and each of these segments is annotated by at least five

independent individuals. This multi-annotator methodology enhances the reliability of the annotations by minimizing individual bias and subjectivity. As part of their assignment, annotators watch the video segments continuously and annotate each one according to the valence/arousal scale. The annotation strategy incorporates both discrete (anger, disgust, fear, happiness, sadness, surprise, and neutral) and continuous (valence and arousal) measures. Comprising 7,371 annotated monologue-based speeches, the OMG dataset serves as a robust platform for researchers exploring the dynamism of emotional behaviors within singular narratives.

4.4 Evaluation

When we delve into deep MER models, we find that their methods of handling continuous emotions and discrete emotions differ. For the prediction of discrete emotions, the model typically employs the cross-entropy loss function, which is a commonly used loss function in classification problems, aiming to minimize the difference between the probability distribution predicted by the model and the probability distribution of the true labels.

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(p_1), \tag{4.1}$$

where N is the number of classes, y_i is the true label, and p_i is the probability predicted by the model for that category.

As for the prediction of continuous emotions, due to their nature learning more toward regression problems, the model tends to use MSE (Mean Squared Error) and MAE (Mean Absolute Error) as loss functions.

$$\text{MSE} = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 \tag{4.2}$$

$$\text{MAE} = \frac{1}{M}\sum_{i=1}^{M}|y_i - \hat{y}_i| \tag{4.3}$$

where M is the total number of samples, y_i is the true continuous emotion value, and $\hat{y}_i$ is

the predicted value by the model.

We summarize two primary types of evaluation metrics for emotion recognition models: one for classification models, denoted as classification evaluation metrics, and another for continuous emotion prediction models, referred to as regression evaluation metrics.

4.4.1 Weighted Average Accuracy (ACC)

The ACC is a widely adopted evaluation metric for emotion classification tasks. It considers the imbalanced distribution of samples across various categories and computes the weighted accuracy for each category based on its sample proportion to acquire the overall accuracy. The calculation formula is as follows

$$\mathrm{ACC} = \frac{1}{N} w_i \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}, \tag{4.4}$$

where w_i is the weight of class i, TP_i is the number of true positive in class i, TN_i is the number of true negative in class i, FP_i is the number of false positive in class i, and FN_i is the number of false negative in class i.

4.4.2 Unweighted Average Accuracy (UACC)

The UACC is calculated by computing the mean of accuracies for different emotion categories while disregarding any imbalances between these categories. The following formula represents the calculation process

$$\mathrm{UACC} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}, \tag{4.5}$$

4.4.3 Weighted Average F1 (F1)

The F1 score is a widely used evaluation metric in the field of emotion classification. It is calculated by combining precision and recall according to the formula provided below:

$$\mathrm{F1} = \frac{1}{N} \sum_{i=1}^{N} w_i \frac{2Precision_i \times Recall_i}{Precision_i + Recall_i}, \tag{4.6}$$

where $Precision_i = \frac{TP_i}{TP_i+FP_i}$ is thje precision of class i, Recall $i = \frac{TP_i}{TP_i+FN_i}$ is the recall of class i.

4.4.4 Unweighted Average F1 (UF1)

The UF1 score computes the arithmetic average of F1 scores for each emotion category. Its calculation formula is shown below

$$\mathrm{UF1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{4.7}$$

4.4.5 Mean Squared Error (MSE)

The MSE is a commonly employed evaluation metric for regression tasks. It calculates the average of the squared differences between the predicted values and the true labels. For continuous emotion prediction tasks, a lower MSE signifies that the model's predictions align more closely with the true emotion values.

4.4.6 Root Mean Squared Error (RMSE)

The RMSE is obtained by calculating the square root of the MSE. A lower RMSE indicates fewer prediction errors in the model. The calculation process is outlined as follows

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{4.8}$$

4.4.7 Pearson Correlation Coefficient (PCC)

The PCC quantifies the linear relationship between the predicted values of a model and the actual emotional values. PCC values fall within the range of -1 to 1, where values near 1 signify a strong positive correlation, values near -1 denote a significant negative correlation, and values near 0 signify no correlation. The PCC is calculated as follows

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N}(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_{i=1}^{M}(y_i - \mu_y)\sum_{i=1}^{M}(\hat{y}_i - \mu_{\hat{y}})^2}}, \tag{4.9}$$

where μ_y and $\mu_{\hat{y}}$ are the means of actual and predicted emotion scores.

4.4.8 Concordance Correlation Coefficient (CCC)

The CCC combines the advantages of PCC and MSE. It not only captures the covariation relationship between predictions and ground truth but also reflects its deviation. Therefore, it provides a better reflection of the alignment between predictions and ground truth, making it a widely utilized performance evaluation metric for continuousdimensional sentiment prediction. Higher CCC values indicate strong performance in terms of consistency and accuracy. The specific calculation process for CCC is as follows:

$$\mathrm{CCC} = \frac{2\rho\,\sigma_y\,\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \tag{4.10}$$

where ρ is the Pearson correlation coefficient and σ_y and $\sigma_{\hat{y}}$ are the standard deviations of actual and predicted emotion scores.

Chapter 5

Graph-based Multimodal Emotion Recognition

Multimodal emotion recognition aims to recognize emotions for each utterance of multiple modalities, which has received increasing attention for its application in human-machine interaction. Graph Neural Networks can well characterize the interaction between different utterances, achieving state-of-the-art performance on many benchmarks.

We first give an overall introduction to current multimodal emotion recognition models in Section 5.1.1. Then, we summary the limitations of current studies in Section 5.1.2. Specifically, we argue that current graph-based methods fail to simultaneously depict global contextual features and local diverse unimodal features in a dialogue. Furthermore, with the number of graph layers increasing, they easily fall into over-smoothing. Finally, we introduce the proposed solutions in Section 5.1.3.

Then, we give a detailed introduction about the necessary background knowledge and the most related works in Section 5.2.

Next, in Section 5.3, we introduce the proposed method: **jo**int modalit**y** **fu**sion and graph contrastive **l**earning for multimodal emotion recognition (JOYFUL), where multimodal fusion, contrastive learning, and emotion recognition are jointly optimized. Specifically, in Section 5.3.3, we first design a new multimodal fusion mechanism that can provide deep interaction and fusion between the global contextual and unimodal specific features. Then, in Section 5.3.4, we introduce a graph contrastive learning framework with inter-view and

intra-view contrastive losses to learn more distinguishable representations for samples with different sentiments. Finally, in Section 5.3.5, we use an multimodal emotion recognition layer for the emotion classification.

Finally, we introduce the experimental settings in Section 5.4 and experimental results in Section 5.5. Extensive experiments on three benchmark datasets indicate that JOYFUL achieved state-of-the-art (SOTA) performance compared to all baselines. [1] We also give the limitations and summary of this study in Section 5.6.

5.1 Introduction

5.1.1 Current Studies for Multimodal Emotion Recognition

"Integration of information from multiple sensory channels is crucial to understand the tendencies and reactions of humans" (Partan and Marler, 1999). Multimodal emotion recognition in conversations (MERC) aims exactly at identifying and tracking the emotional state of each utterance from heterogeneous visual, audio, and text channels. Due to its potential applications in the creation of human-computer interaction systems (Li et al., 2022c), social media analysis (Gupta et al., 2022; Wang et al., 2023), and recommendation systems (Singh et al., 2022), MERC has received increasing attention in the natural language processing (NLP) community (Poria et al., 2019b, 2021), which even has the potential to be widely applied in other tasks such as question answering (Ossowski and Hu, 2023; Wang et al., 2022c), text generation (Liang et al., 2023; Zhang et al., 2023; Li et al., 2022a) and bioinformatics (Nicolson et al., 2023; You et al., 2022).

Figure 5.1 shows that emotions expressed in a dialogue are affected by three main factors: 1) multiple uni-modalities, *e.g.*, different modalities complete each other to provide a more informative utterance representation; 2) global contextual information, *e.g.*, u_3^A depends on the topic "The ship sank into the sea", indicating fear; 3) intra-person and inter-person dependencies, *e.g.*, u_6^A becomes sad affected by sadness in u_4^B and u_5^B.

Depending on how to model intra-person and inter-person dependencies, current MERC

[1] Code is released on Github (https://anonymous/MERC).

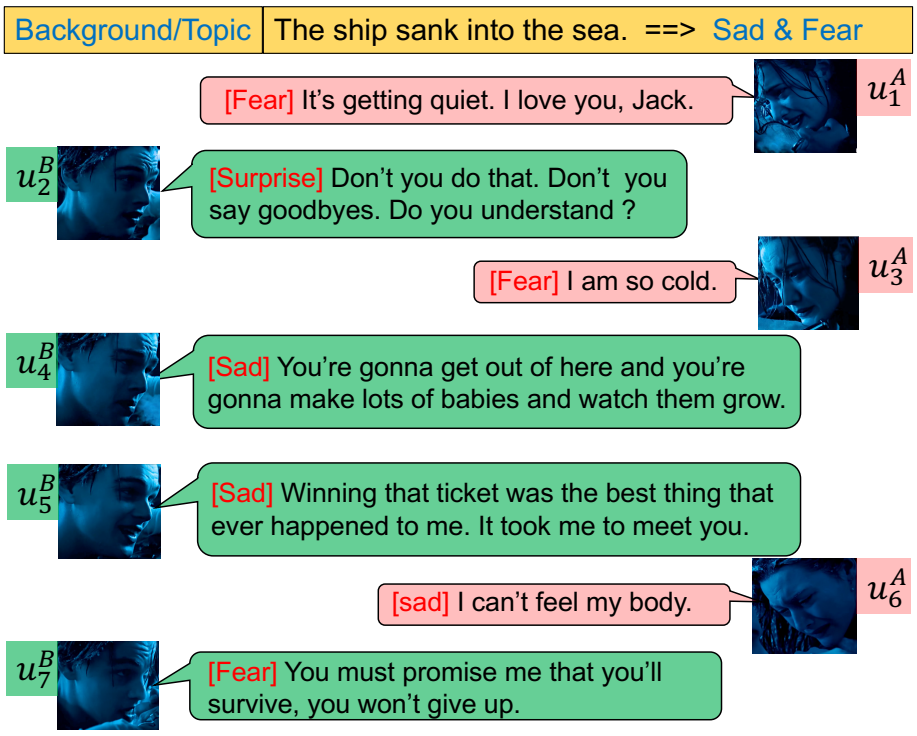


Figure 5.1: Emotions are affected by multiple modalities, global contextual, intra- and inter-person dependencies. Images are from the movie "Titanic".

methods can be categorized into Sequence-based and Graph-based methods. The former (Dai et al., 2021; Mao et al., 2022; Liang et al., 2022) use recurrent neural networks or Transformers to model the temporal interaction between utterances. However, they failed to distinguish intra-speaker and inter-speaker dependencies and easily lost unimodal specific features by the cross-modal attention mechanism (Rajan et al., 2022). Graph structure (Joshi et al., 2022; Wei et al., 2019) solves these issues by using edges between nodes (speakers) to distinguish intra-speaker and inter-speaker dependencies. Graph Neural Networks (GNNs) further help nodes learn common features by aggregating information from neighbors while maintaining their unimodal specific features.

5.1.2 Limitations of Previous Studies

Although graph-based MERC methods have achieved great success, there still remain problems that need to be solved:

(1) Current methods directly aggregate features of multiple modalities (Joshi et al., 2022) or project modalities into a latent space to learn representations (Li et al., 2022f), which ignores the diversity of each modality and fails to capture richer semantic information from each modality. They also ignore global contextual information during the feature fusion process, leading to poor performance.

(2) Since all graph-based methods adopt GNN (Scarselli et al., 2009) or Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017), with the number of layers deepening, the phenomenon of over-smoothing starts to appear, resulting in the representation of similar sentiments being indistinguishable.

(3) Most methods use a two-phase pipeline (Fu et al., 2021; Joshi et al., 2022), where they first extract and fuse unimodal features as utterance representations and then fix them as input for graph models. However, the two-phase pipeline will lead to sub-optimal performance since the fused representations are fixed and cannot be further improved to benefit from the downstream supervisory signals.

5.1.3 Proposed Solutions

To solve the above-mentioned problems, we propose **Jo**int multimodalit**y** **fu**sion and graph contrastive **l**earning for MERC (JOYFUL), where multimodal fusion, graph contrastive learning (GCL), and multimodal emotion recognition are jointly optimized in an overall objective function. Specifically, we show the detailed processes as follows:

(1) We first design a new multimodal fusion mechanism that can simultaneously learn and fuse a global contextual representation and unimodal specific representations. For the global contextual representation, we smooth it with a proposed topic-related vector to maintain its consistency, where the topic-related vector is temporally updated. For unimodal specific representations, we project them into a shared subspace to fully explore their richer semantics without losing alignment with other modalities.

(2) To alleviate the over-smoothing issue of deeper GNN layers, which showed that contrastive learning could provide more distinguishable node representations to benefit various downstream tasks, we propose a GCL-based cross-view framework to alleviate the difficulty of categorizing similar emotions, which helps to learn more distinctive representations of utterance by making samples with the same sentiment cohesive and those with different sentiments mutually exclusive. Furthermore, graph augmentation strategies are designed to improve JOYFUL's robustness and generalizability.

(3) We jointly optimize each part of JOYFUL in a *end-to-end* manner to ensure optimized global performance. Extensive experiments conducted on three multimodal benchmark datasets demonstrated the effectiveness and robustness of JOYFUL.

5.2 Related Work and Background Knowledge

In this section, we introduce the most related work and background knowledge of our proposed methods. Specifically, we give a detailed introduction to the most related multimodal emotion recognition models in Section 5.2.1. Then, we introduce the classic multimodal fusion mechanisms in Section 5.2.2. Finally, we introduce the graph contrastive learning strategies in Section 5.2.3.

5.2.1 Multimodal Emotion Recognition

Depending on how to model the context of utterances, existing MERC methods are categorized into three classes: Recurrent-based methods (Majumder et al., 2019; Mao et al., 2022) adopt RNN or LSTM to model the sequential context for each utterance. Transformers-based methods (Ling et al., 2022; Liang et al., 2022; Le et al., 2022) use Transformers with cross-modal attention to model the intra- and inter-speaker dependencies. Graph-based methods (Joshi et al., 2022; Zhang et al., 2021a; Fu et al., 2021) can control context information for each utterance and provide accurate intra- and inter-speaker dependencies, achieving SOTA performance on many MERC benchmark datasets.

5.2.2 Multimodal Fusion Mechanism

Learning effective fusion mechanisms is one of the key challenges in multimodal learning (Shankar, 2022). By capturing the interactions between different modalities in a more reasonable way, deep models can acquire more comprehensive information. Current fusion methods can be classified into aggregation-based (Wu et al., 2021; Guo et al., 2021), alignment-based (Liu et al., 2020; Li et al., 2022f), and their mixture (Wei et al., 2019; Nagrani et al., 2021). Aggregation-based fusion methods (Zadeh et al., 2017; Chen et al., 2021) adopt concatenation, tensor fusion, and memory fusion to combine multiple modalities. Alignment-based fusion is based on latent cross-modal adaptation, which adapts the flows from one modality to another (Wang et al., 2022a). Different from the above methods, we learn global contextual information by concatenation while fully exploring the specific patterns of each modality in an alignment manner.

5.2.3 Graph Contrastive Learning

GCL aims to learn representations by maximizing feature consistency in differently augmented views, that exploit data- or task-specific augmentations, to inject the desired feature invariance (You et al., 2020). GCL has been well used in the NLP community through self-supervised and supervised settings. Self-supervised GCL first creates augmented graphs by edge/node deletion and insertion (Zeng and Xie, 2021), or attribute masking (Zhang et al., 2022c). It then captures the intrinsic patterns and properties in the augmented graphs without using human-provided labels. Supervised GCL designs adversarial (Sun et al., 2022) or geometric (Li et al., 2022e) contrastive loss to make full use of the information on the label. For example, Li et al. (2022d) first used supervised CL for emotion recognition, greatly improving performance. Inspired by previous studies, we jointly consider self-supervised (suitable graph augmentation) and supervised (cross-entropy) manners to fully explore graph structural information and downstream supervisory signals.

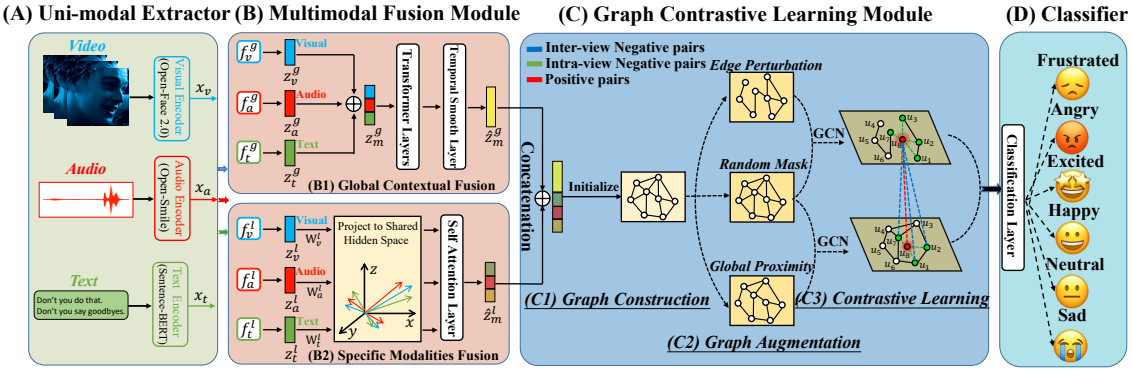


Figure 5.2: Overview of JOYFUL. We first extract unimodal features, fuse them using a multimodal fusion module, and use them as input of the GCL-based framework to learn better representations for emotion recognition.

5.3 Methodology: JOYFUL

Figure 5.2 shows an overview of JOYFUL, which consists mainly of four components: (A) a *unimodal extractor* introduced in Section 5.3.2, (B) a *multimodal fusion* (MF) module introduced in Section 5.3.3, (C) a *graph contrastive learning* module introduced in Section 5.3.4, and (D) a *classifier* introduced in Section 5.3.5. We first give a formal notation and the task definition of JOYFUL, and then introduce each component in detail.

5.3.1 Notations and Task Definition

In dialogue emotion recognition, a training dataset $\mathcal{D} = \{(\mathcal{C}_i, \mathcal{Y}_i)\}_{i=1}^{N}$ is given, where $\mathcal{C}_i$ represents the i-th conversation, each conversation contains several utterances $\mathcal{C}_i = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$, and $\mathcal{Y}_i \in \mathbf{Y}^m$, given label set $\mathbf{Y} = \{y_1, \ldots, y_k\}$ of k emotion classes. Let $\mathbf{X}^v$, $\mathbf{X}^a$, $\mathbf{X}^t$ be the visual, audio, and text feature spaces, respectively. The goal of MERC is to learn a function $\mathbf{F} : \mathbf{X}^v \times \mathbf{X}^a \times \mathbf{X}^t \rightarrow \mathbf{Y}$ that can recognize the emotion label for each utterance. We used three widely used multimodal conversational benchmark datasets, namely IEMOCAP, MOSEI, and MELD, to evaluate the performance of our model. Please refer to Section 5.4.1 for their detailed statistical information.

5.3.2 Unimodal Extractor

For IEMOCAP (Busso et al., 2008), video features $\boldsymbol{x}_v \in \mathbb{R}^{512}$, audio features $\boldsymbol{x}_a \in \mathbb{R}^{100}$, and text features $\boldsymbol{x}_t \in \mathbb{R}^{768}$ are obtained from OpenFace (Baltrusaitis et al., 2018), OpenSmile (Eyben et al., 2010) and SBERT (Reimers and Gurevych, 2019), respectively. For MELD (Poria et al., 2019a), $\boldsymbol{x}_v \in \mathbb{R}^{342}$, $\boldsymbol{x}_a \in \mathbb{R}^{300}$, and $\boldsymbol{x}_t \in \mathbb{R}^{768}$ are obtained from DenseNet (Huang et al., 2017a), OpenSmile, and TextCNN (Kim, 2014). For MOSEI (Zadeh et al., 2018b), $\boldsymbol{x}_v \in \mathbb{R}^{35}$, $\boldsymbol{x}_a \in \mathbb{R}^{80}$, and $\boldsymbol{x}_t \in \mathbb{R}^{768}$ are obtained from TBJE (Delbrouck et al., 2020), LibROSA (Raguraman et al., 2019), and SBERT. Textual features are sentence-level static features. Audio and visual modalities are utterance-level features by averaging all the token features.

5.3.3 Multimodal Fusion Module

Though the unimodal extractors can capture long-term temporal context, they are unable to handle feature redundancy and noise due to the modality gap. Thus, we design a new multimodal fusion module (Figure 5.2 (B)) to inherently separate multiple modalities into two disjoint parts, contextual representations and specific representations, to extract the consistency and specificity of heterogeneous modalities collaboratively and individually.

5.3.3.1 Contextual Representation Learning

Contextual representation learning aims to explore and learn hidden contextual intent/topic knowledge of the dialogue, which can greatly improve the performance of JOYFUL. In Figure 5.2 (B1), we first project all unimodal inputs $\boldsymbol{x}_{\{v,a,t\}}$ into a latent space by using three separate connected deep neural networks $f^g_{\{v,a,t\}}(\cdot)$ to obtain hidden representations $\boldsymbol{z}^g_{\{v,a,t\}}$. Then, we concatenate them as $\boldsymbol{z}^g_m$ and apply it to a multi-layer transformer to maximize the correlation between multimodal features, where we learn a global contextual multimodal representation $\hat{\boldsymbol{z}}^g_m$. Considering that contextual information will change over time, we design a temporal smoothing strategy for $\hat{\boldsymbol{z}}^g_m$ as

$$\mathcal{J}_{smooth} = \|\hat{\boldsymbol{z}}^g_m - \boldsymbol{z}_{con}\|^2, \tag{5.1}$$

where $\boldsymbol{z}_{con}$ is the topic-related vector describing the high-level global contextual information without requiring topic-related inputs, following the definition in Joshi et al. (2022). We update the (i+1)-th utterance as follows:

$$\boldsymbol{z}_{con} \leftarrow \boldsymbol{z}_{con} + e^{\eta * i} \hat{\boldsymbol{z}}_m^g, \tag{5.2}$$

where η is the exponential smoothing parameter (Shazeer and Stern, 2018), indicating that more recent information will be more important.

To ensure fused contextual representations capture enough details from hidden layers, Hazarika et al. (2020) minimized the reconstruction error between fused representations with hidden representations. Inspired by their work, to ensure that $\hat{\boldsymbol{z}}_m^g$ contains essential modality cues for downstream emotion recognition, we reconstruct $\boldsymbol{z}_m^g$ from $\hat{\boldsymbol{z}}_m^g$ by minimizing their Euclidean distance:

$$\mathcal{J}_{rec}^g = \|\hat{\boldsymbol{z}}_m^g - \boldsymbol{z}_m^g\|^2. \tag{5.3}$$

5.3.3.2 Specific Representation Learning

Specific representation learning aims to fully explore specific information from each modality to complement one another. Figure 5.2 (B2) shows that we first use three fully connected deep neural networks $f_{\{v,a,t\}}^{\ell}(\cdot)$ to project unimodal embeddings $\boldsymbol{x}_{\{v,a,t\}}$ into a hidden space with representations as $\boldsymbol{z}_{\{v,a,t\}}^{\ell}$. Considering that visual, audio, and text features are extracted with different encoding methods, directly applying multiple specific features as an input for the downstream emotion recognition task will degrade the model's accuracy. To solve it, the multimodal features are projected into a shared subspace, and a shared trainable basis matrix is designed to learn aligned representations for them. Therefore, the multimodal features can be fully integrated and interacted to mitigate feature discontinuity and remove noise across modalities. We define a shared trainable basis matrix $\mathbf{B}$ with q basis vectors as $\mathbf{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q)^T \in \mathbb{R}^{q \times d_b}$ with d_b representing the dimensionality of each basis vector. Here, T indicates transposition. Then, $\boldsymbol{z}_{\{v,a,t\}}^{\ell}$ and $\mathbf{B}$ are projected into the shared subspace:

$$\tilde{\boldsymbol{z}}_{\{v,a,t\}}^{\ell} = \mathbf{W}_{\{v,a,t\}} \boldsymbol{z}_{\{v,a,t\}}^{\ell}, \quad \widetilde{\mathbf{B}} = \mathbf{B}\mathbf{W}_b, \tag{5.4}$$

where $\mathbf{W}_{\{v,a,t,b\}}$ are trainable parameters. To learn new representations for each modality, we calculate the cosine similarity between them and $\mathbf{B}$ as

$$S_{ij}^{\{v,a,t\}} = (\tilde{\boldsymbol{z}}_{\{v,a,t\}}^{\ell})_i \cdot \widetilde{\boldsymbol{b}}_j, \tag{5.5}$$

where S_{ij}^{v} denotes the similarity between the i-th visual feature $(\tilde{\boldsymbol{z}}_{v}^{\ell})_i$ and the j-th basis vector representation $\widetilde{\boldsymbol{b}}_j$. To prevent inaccurate representation learning caused by an excessive weight of a certain item, the similarities are further normalized by

$$S_{ij}^{\{v,a,t\}} = \frac{\exp\left(S_{ij}^{\{v,a,t\}}\right)}{\sum_{k=1}^{q} \exp\left(S_{ik}^{\{v,a,t\}}\right)}. \tag{5.6}$$

Then, the new representations are obtained as

$$(\hat{\boldsymbol{z}}_{\{v,a,t\}}^{\ell})_i = \sum_{k=1}^{q} S_{ik}^{\{v,a,t\}} \cdot \widetilde{\boldsymbol{b}}_k, \tag{5.7}$$

where $\hat{\boldsymbol{z}}_{\{v,a,t\}}^{\ell}$ are new representations, and we use reconstruction loss for the combinations

$$\mathcal{J}_{rec}^{\ell} = \|\hat{\boldsymbol{z}}_{m}^{\ell} - \boldsymbol{z}_{m}^{\ell}\|^2, \tag{5.8}$$

where *Concat(,)* indicates the concatenation operation, i.e.,

$$\hat{\boldsymbol{z}}_{m}^{\ell} = \mathit{Concat}(\hat{\boldsymbol{z}}_{v}^{\ell}, \hat{\boldsymbol{z}}_{a}^{\ell}, \hat{\boldsymbol{z}}_{t}^{\ell}), \quad \boldsymbol{z}_{m}^{\ell} = \mathit{Concat}(\boldsymbol{z}_{v}^{\ell}, \boldsymbol{z}_{a}^{\ell}, \boldsymbol{z}_{t}^{\ell}) \tag{5.9}$$

Finally, we define the multimodal fusion loss by combining Eqs.(5.1), (5.3), and (5.8) as

$$\mathcal{L}_{mf} = \mathcal{J}_{smooth} + \mathcal{J}_{rec}^{g} + \mathcal{J}_{rec}^{\ell}. \tag{5.10}$$

5.3.4 Graph Contrastive Learning Module

5.3.4.1 Graph Construction

Graph construction aims to establish relations between past and future utterances that preserve both intra- and inter-speaker dependencies in a dialogue. We define the i-th

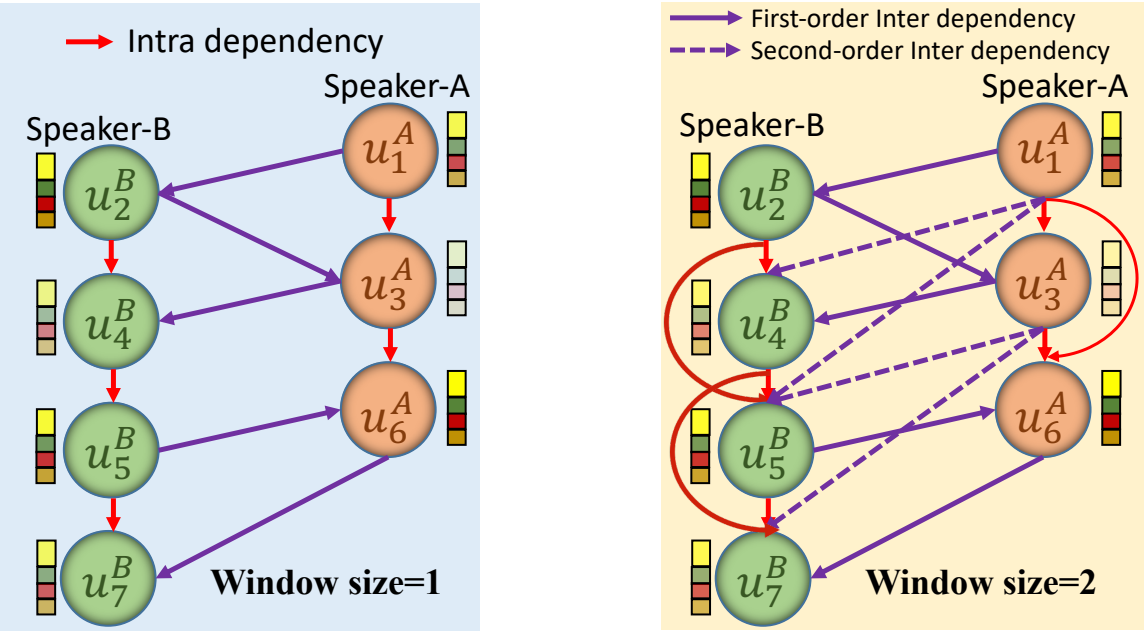


Figure 5.3: An example of graph construction.

dialogue with P speakers as $\mathcal{C}_i = \{\mathcal{U}^{S_1}, \ldots, \mathcal{U}^{S_P}\}$, where $\mathcal{U}^{S_i} = \{\boldsymbol{u}_1^{S_i}, \ldots, \boldsymbol{u}_m^{S_i}\}$ represents the set of utterances spoken by speaker S_i. Following Ghosal et al. (2019), we define a graph with nodes representing utterances and directed edges representing their relations: $\mathcal{R}_{ij} = \boldsymbol{u}_i \rightarrow \boldsymbol{u}_j$, where the arrow represents the speaking order. *Intra-Dependency* ($\mathcal{R}_{intra} \in \{\mathcal{U}^{S_i} \rightarrow \mathcal{U}^{S_i}\}$) represents intra-relations between the utterances (red lines), and *Inter-Dependency* ($\mathcal{R}_{inter} \in \{\mathcal{U}^{S_i} \rightarrow \mathcal{U}^{S_j}\}, i \neq j$) represents the inter-relations between the utterances (purple lines), as shown in Figure 5.3. All nodes are initialized by concatenating contextual and specific representations as $\boldsymbol{h}_m = Concat(\hat{\boldsymbol{z}}_m^g, \hat{\boldsymbol{z}}_m^\ell)$. And we show that window size is a hyper-parameter that controls the context information for each utterance and provide accurate intra- and inter-speaker dependencies.

5.3.4.2 Graph Augmentation

Graph Augmentation (GA): Inspired by Zhu et al. (2020), creating two augmented views by using different ways to corrupt the original graph can provide highly heterogeneous contexts for nodes. By maximizing the mutual information between two augmented views,

we can improve the robustness of the model and obtain distinguishable node representations. However, there are no universally appropriate GA methods for various downstream tasks (Xu et al., 2021), which motivates us to design specific GA strategies for MERC. Considering that MERC is sensitive to initialized representations of utterances, intra-speaker and inter-speaker dependencies, we design three corresponding GA methods:

(1) **Feature Masking** (FM): given the initialized representations of utterances, we randomly select p dimensions of the initialized representations and mask their elements with zero, which is expected to enhance the robustness of JOYFUL to multimodal feature variations;

(2) **Edge Perturbation** (EP): given the graph $\mathcal{G}$, we randomly drop and add $p\%$ of intra- and inter-speaker edges, which is expected to enhance the robustness of JOYFUL to local structural variations;

(3) **Global Proximity** (GP): given the graph $\mathcal{G}$, we first use the Katz index (Katz, 1953) to calculate high-order similarity between intra-speakers and inter-speakers, and randomly add $p\%$ high-order edges between speakers, which is expected to enhance the robustness of JOYFUL to global structural variations. To make it easier to understand, we give an example of global proximity. Specifically, as shown in Figure 5.4, given the network $\mathcal{G}$ and a modified p, we first used the Katz index (Katz, 1953) to calculate a high-order similarity between the vertices. We considered an arbitrarily large number of high-order distances. For example, second-order similarity between u_1^A and u_4^B as $u_1^A \rightarrow u_4^B = 0.83$, third-order similarity between u_1^A and u_5^B as $u_1^A \rightarrow u_5^B = 0.63$, and fourth-order similarity between u_1^A and u_7^B as $u_1^A \rightarrow u_7^B = 0.21$. We then define the threshold score as 0.5, where a high-order similarity score less than the threshold will not be selected as added edges. Finally, we randomly selected $p\%$ edges and added them to the original graph $\mathcal{G}$ to construct the augmented graph.

We propose a hybrid scheme for generating graph views on both structure and attribute levels to provide diverse node contexts for the contrastive objective. Figure 5.2 (C) shows that the combination of (FM $\&$ EP) and (FM $\&$ GP) are used to obtain two correlated views.

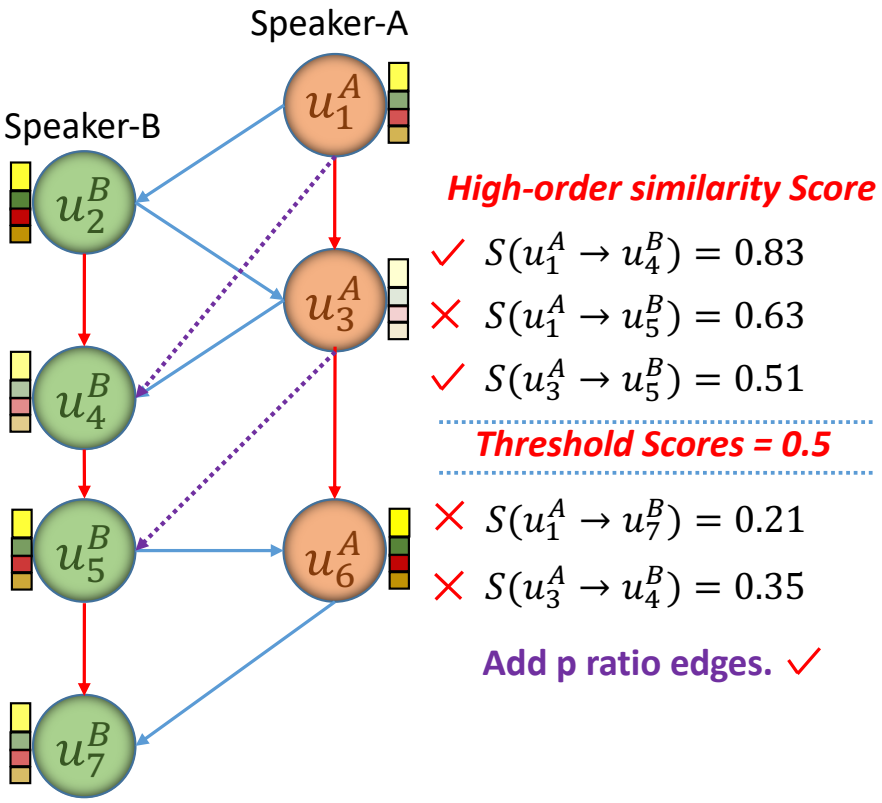


Figure 5.4: Example of adding $p\%$ high-order edges to explore global topological information of graph.

5.3.4.3 Graph Contrastive Learning

Graph contrastive learning adopts an L-th layer GCNs as a graph encoder to extract node hidden representations $\mathbf{H}^{(1)} = \{\boldsymbol{h}_1^{(1)}, \ldots, \boldsymbol{h}_m^{(1)}\}$ and $\mathbf{H}^{(2)} = \{\boldsymbol{h}_1^{(2)}, \ldots, \boldsymbol{h}_m^{(2)}\}$ for two augmented graphs, where $\boldsymbol{h}_i$ is the hidden representation for the i-th node. We follow an iterative neighborhood aggregation (or message passing) scheme to capture the structural information within the nodes' neighborhood. Formally, the propagation and aggregation of the ℓ-th GCN layer is:

$$\boldsymbol{a}_{(i,\ell)} = \text{AGG}_{(\ell)}\left(\{\boldsymbol{h}_{(j,\ell-1)} | j \in \mathbf{N}_i\}\right) \tag{5.11}$$

$$\boldsymbol{h}_{(i,\ell)} = \text{COM}_{(\ell)}\left(\boldsymbol{h}_{(i,\ell-1)} \oplus \boldsymbol{a}_{(i,\ell)}\right), \tag{5.12}$$

where $\boldsymbol{h}_{(i,\ell)}$ is the embedding of the i-th node at the ℓ-th layer, $\boldsymbol{h}_{(i,0)}$ is the initialization of the i-th utterance, $\mathbf{N}_i$ represents all neighbour nodes of the i-th node, and $\text{AGG}_{(\ell)}(\cdot)$ and $\text{COM}_{(\ell)}(\cdot)$ are aggregation and combination of the ℓ-th GCN layer (Hamilton et al., 2017). For convenience, we define $\boldsymbol{h}_i = \boldsymbol{h}_{(i,L)}$. After the L-th GCN layer, final node representations of two views are $\mathbf{H}^{(1)}$ / $\mathbf{H}^{(2)}$.

In Figure 5.2 (C3), we design the intra- and inter-view graph contrastive losses to learn distinctive node representations. We start with the inter-view contrastiveness, which pulls closer the representations of the same nodes in two augmented views while pushing other nodes away, as depicted by the red and blue dash lines in Figure 5.2 (C3). Given the definition of our positive and negative pairs as $(\boldsymbol{h}_i^{(1)}, \boldsymbol{h}_i^{(2)})^+$ and $(\boldsymbol{h}_i^{(1)}, \boldsymbol{h}_j^{(2)})^-$, where $i \neq j$, the inter-view loss for the i-th node is formulated as:

$$\mathcal{L}_{inter}^{i} = -\log \frac{\exp(\text{sim}(\boldsymbol{h}_i^{(1)}, \boldsymbol{h}_i^{(2)}))}{\sum_{j=1}^{m} \exp(\text{sim}(\boldsymbol{h}_i^{(1)}, \boldsymbol{h}_j^{(2)}))}, \tag{5.13}$$

where $\text{sim}(\cdot,\cdot)$ denotes the similarity between two vectors, i.e., the cosine similarity.

Intra-view contrastiveness regards all nodes except the anchor node as negatives within a particular view (green dash lines in Figure 5.2 (C3)), as defined $(\boldsymbol{h}_i^{(1)}, \boldsymbol{h}_j^{(1)})^-$ where $i \neq j$. The intra-view contrastive loss for the i-th node is defined as:

$$\mathcal{L}_{intra}^{i} = -\log \frac{\exp(\text{sim}(\boldsymbol{h}_i^{(1)}, \boldsymbol{h}_i^{(2)}))}{\sum_{j=1}^{m} \exp(\text{sim}(\boldsymbol{h}_i^{(1)}, \boldsymbol{h}_j^{(1)}))}. \tag{5.14}$$

By combining the inter- and intra-view contrastive losses of Eqs.(5.13) and (5.14), the contrastive objective function $\mathcal{L}_{ct}$ is formulated as:

$$\mathcal{L}_{ct} = \frac{1}{2m} \sum_{i=1}^{m} (\mathcal{L}_{inter}^{i} + \mathcal{L}_{intra}^{i}). \tag{5.15}$$

Algorithm 2: Overall process of JOYFUL

input : Visual features $\boldsymbol{x}_v$;
Audio features $\boldsymbol{x}_a$;
Text features $\boldsymbol{x}_t$;
Parameters: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, Window size

output : Emotion recognition label.

Initialize trainable parameters;
for $epoch \leftarrow 1$ **to** $epoch\ num$ **do**
  Global Contextual Fusion $\hat{\boldsymbol{z}}_m^g$;
  Specific Modality Fusion $\hat{\boldsymbol{z}}_m^\ell = (\boldsymbol{z}_v^g \| \boldsymbol{z}_a^g \| \boldsymbol{z}_t^g)$;
  // **Compute multimodal fusion loss**
  Compute $\mathcal{L}_{mf}$, in accordance with Eq.(5.10);
  Feature Concatenation $\boldsymbol{h} = (\hat{\boldsymbol{z}}_m^g \| \hat{\boldsymbol{z}}_m^\ell)$;
  Adopt $\boldsymbol{h}$ as initialization for Graph;
  // **Generate two augmented views**
  Apply FM & EP to generate view: $\mathcal{G}^{(1)}$;
  Apply FM & GP to generate view: $\mathcal{G}^{(2)}$;
  // **Extract features of two views**
  $\mathbf{H}^{(1)} = GCNs(\mathcal{G}^{(1)})$, $\mathbf{H}^{(2)} = GCNs(\mathcal{G}^{(2)})$;
  // **Compute contrastive learning loss**
  Compute $\mathcal{L}_{ct}$, in accordance with Eq.(5.15) ;
  // **Aggregate extracted features**
  $\mathbf{H} = \mathbf{H}^{(1)} + \mathbf{H}^{(2)}$;
  // **Compute emotion recognition loss**
  Compute $\mathcal{L}_{ce}$, in accordance with Eq.(6.5);
  // **Joint training**
  Compute $\mathcal{L}_{all}$, in accordance with Eq.(5.17);
  // **Optimize with Adam optimizer**
end
Adopt classifier on $\mathbf{H}$ to predict the emotional label.

5.3.5 Emotion Recognition Classifier

We use cross-entropy loss for classification as:

$$\mathcal{L}_{ce} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} y_i^j \log(\hat{y}_i^j), \tag{5.16}$$

where k is the number of emotion classes, m is the number of utterances, $\hat{y}_i^j$ is the i-th predicted label, and y_i^j is the i-th ground truth of j-th class.

Above all, combining the MF loss of Eq.(5.10), contrastive loss of Eq.(5.15), and classification loss of Eq.(6.5) together, the final objective function is

$$\mathcal{L}_{all} = \boldsymbol{\alpha}\mathcal{L}_{mf} + \boldsymbol{\beta}\mathcal{L}_{ct} + \mathcal{L}_{ce}, \tag{5.17}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the trade-off hyper-parameters. We give pseudocode in Algorithm 2.

Dataset	Train	Valid	Test
IEMOCAP(4-way)	3,200/108	400/12	943/31
IEMOCAP(6-way)	5,146/108	664/12	1,623/31
MELD	9,989/1,039	1,109/114	2,80/2,610
MOSEI	16,327/2,249	1,871/300	4,662/679

Table 5.1: Utterances/Conversations of four datasets.

5.4 Experimental Settings

5.4.1 Datasets and Metrics

In Table 5.1, IEMOCAP is a conversational dataset where each utterance was labeled with one of the six emotion categories (Anger, Excited, Sadness, Happiness, Frustrated and Neutral). Following COGMEN, two IEMOCAP settings were used for testing, one with four emotions (Anger, Sadness, Happiness, and Neutral) and one with all six emotions, where 4-way directly removes the additional two emotion labels (Excited and Frustrated). MOSEI was labeled with six emotion labels (Anger, Disgust, Fear, Happiness, Sadness,

and Surprise). For six emotion labels, we performed two settings: *binary classification* considers the target emotion as one class and all other emotions as another class, and *multi-label classification* tags multiple labels for each utterance. MELD was labeled with six universal emotions (Joy, Sadness, Fear, Anger, Surprise, and Disgust). We split the datasets into 70%/10%/20% as training/validation/test data, respectively. Following Joshi et al. (2022), we used *Accuracy* and *Weighted F1-score* (WF1) as evaluation metrics. And we list the detailed label distribution of MELD in Table 5.2, IEMOCAP 4-way in Table 5.3, IEMOCAP 6-way in Table 5.4 and MOSEI in Table 5.5.

MELD	Train	Valid	Test
Anger	1,109	153	345
Disgusted	271	22	68
Fear	268	40	50
Joy	1,743	163	402
Neutral	4,710	470	1,256
Sadness	683	111	208
Surprise	1,205	150	281
Total	9,989	1,109	2,610

Table 5.2: Labels distribution of MELD dataset.

IEMOCAP 4-way	Train	Valid	Test
Happy	453	51	144
Sad	783	56	245
Neutral	1,092	232	384
Angry	872	61	170
Total	3,200	400	943

Table 5.3: Labels distribution of IEMOCAP 4-way.

5.4.2 Implementation Details

We selected the augmentation pairs (FM & EP) and (FM & GP) for two views. We set the augmentation ratio p=20% and smoothing parameter η=0.2, and applied the Adam (Kingma

IEMOCAP 6-way	Train	Valid	Test
Happy	459	45	144
Sad	746	93	245
Neutral	1,161	163	384
Angry	854	79	170
Excited	576	166	299
Frustrated	1,350	118	381
Total	5,146	644	1,623

Table 5.4: Labels distribution of IEMOCAP 6-way.

MOSEI	Train	Valid	Test
Happy	8,735	1,005	2,505
Sad	4,269	520	1,129
Angry	3,526	338	1,071
Surprise	1,642	203	441
Disgusted	2,955	281	805
Fear	1,331	176	385
Total	22,458	2,523	6,336

Table 5.5: Labels distribution of MOSEI dataset.

and Ba, 2015) optimizer with an initial learning rate of $3e$-5. For a fair comparison, we followed the default parameter settings of the baselines and repeated all experiments ten times to report the average accuracy. We list all dimensions of the mathematical symbols of IEMOCAP in Table 5.6. Mathematical symbols for other datasets see our source code.

We conducted the significance test by t-test with Benjamini-Hochberg (*B-H*) (Benjamini and Hochberg, 1995), which is a powerful tool that decreases the false discovery rate. Considering the reproducibility of the multiple significant test, we introduce how we adopt the *B-H* correction and give the hyper-parameter values that we used. Specifically, we first performed a t-test (Yang et al., 1999) with default parameters[2] to calculate the p-value between each comparison method with JOYFUL. We then put the individual p-values in ascending order as input to calculate the p-value corrected using the *B-H* correction.

[2] scipy.stats.ttest_ind.html

Symbols	***Description***
$\boldsymbol{x}_v \in \mathbb{R}^{512}$	Video Features
$\boldsymbol{x}_a \in \mathbb{R}^{100}$	Audio Features
$\boldsymbol{x}_t \in \mathbb{R}^{768}$	Text Features
Contextual Representation Learning	
$\boldsymbol{z}_v^g \in \mathbb{R}^{512}$	Global Hidden Video Features
$\boldsymbol{z}_a^g \in \mathbb{R}^{100}$	Global Hidden Audio Features
$\boldsymbol{z}_t^g \in \mathbb{R}^{768}$	Global Hidden Text Features
$\boldsymbol{z}_m^g \in \mathbb{R}^{1,380}$	Global Combined Features
$\boldsymbol{z}_{con} \in \mathbb{R}^{1,380}$	Topic-related Vector
$\hat{\boldsymbol{z}}_m^g \in \mathbb{R}^{1,380}$	Global Output Features
Specific Representation Learning	
$\boldsymbol{z}_v^\ell \in \mathbb{R}^{460}$	Local Hidden Video Features
$\boldsymbol{z}_a^\ell \in \mathbb{R}^{460}$	Local Hidden Audio Features
$\boldsymbol{z}_t^\ell \in \mathbb{R}^{460}$	Local Hidden Text Features
$\boldsymbol{b}_m \in \mathbb{R}^{460}$	Basic Features
$\tilde{\boldsymbol{z}}_{\{v,a,t\}}^\ell \in \mathbb{R}^{460}$	Features in Shared Space
$\tilde{\boldsymbol{b}}_m \in \mathbb{R}^{460}$	Basic Features in Shared Space
$\mathbf{W}_{\{v,a,t,b\}} \in \mathbb{R}^{460\times460}$	Trainable Matrices
$\hat{\boldsymbol{z}}_{\{v,a,t\}}^\ell \in \mathbb{R}^{460}$	New Multimodal Features
$\hat{\boldsymbol{z}}_m^\ell \in \mathbb{R}^{1,380}$	New Multimodal Combined Features
$\boldsymbol{z}_m^\ell \in \mathbb{R}^{1380}$	Original Combined Features
Graph Contrastive Learning (One GCN Layer)	
$(\hat{\boldsymbol{z}}_g^\ell \| \hat{\boldsymbol{z}}_m^\ell) \in \mathbb{R}^{2,760}$	Global-Local Combined Features
AGG $\in \mathbb{R}^{2,760\times2,760}$	Parameters of Aggregation Layer
COM $\in \mathbb{R}^{2,760\times5,520}$	Input/Output of Combination Layer
$\mathbf{W}_{graph} \in \mathbb{R}^{5,520\times2,760}$	Dimention Reduction after COM
$\boldsymbol{h}_m \in \mathbb{R}^{2,760}$	Node Features of GCN Layer

Table 5.6: Mathematical symbols for IEMOCAP dataset.

We directly use the "***multipletests(*args)***" function from python package[3] and set the hyperparameter of the false discovery rate $Q = 0.05$, which is a widely used default value (Puoliväli et al., 2020). Finally, we obtain a cut-off value as the output of the

[3] statsmodels.stats.multitest.multipletests.html

multipletests function, where cut-off is a dividing line that distinguishes whether two groups of data are significant. If the p-value is smaller than the cut-off value, we can conclude that two groups of data are significantly different.

The use of t-test for testing statistical significance may not be appropriate for F-scores, as mentioned in Dror et al. (2018), as we cannot assume normality. To verify whether our data meet the normality assumption and the homogeneity of variances required for the t-test, following Shapiro and Wilk (1965) and Levene et al. (1960), we conducted the following validation. First, we performed the Shapiro-Wilk test on each group of experimental results to determine whether they are normally distributed. Under the constraint of a significance level (alpha=0.05), all p-values resulting from the Shapiro-Wilk test [4] for the baselines and our model were greater than 0.05. This indicates that the results of the baselines and our model all adhere to the assumption of normality. For example, in IEMOCAP-4, p-values for [Mult, RAVEN, MTAG, PMR, MICA, COGMEN, JOYFUL] are [0.903, 0.957, 0.858, 0.978, 0.970, 0.969, 0.862]. Furthermore, we used the Levene's test (Schultz, 1985) to check for homogeneity of variances between baselines and our model. Under the constraint of a significance level (alpha = 0.05), we found that our p-values are greater than 0.05, indicating the homogeneity of the variances between the baselines and our model. For example, we obtained p-values 0.3101 and 0.3848 for group-based baselines on IEMOCAP-4 and IEMOCAP-6, respectively. Since we were able to demonstrate that all baselines and our model conform to the assumptions of normality and homogeneity of variances, we believe that the significance tests we reported are accurate.

5.4.3 Baselines

Different MERC datasets have different best system results, we selected SOTA baselines for each dataset. For IEMOCAP-4, we selected Mult (Tsai et al., 2019a), RAVEN (Wang et al., 2019), MTAG (Yang et al., 2021), PMR (Lv et al., 2021), COGMEN and MICA (Liang et al., 2021) as our baselines. For IEMOCAP-6, we selected Mult, FE2E (Dai et al., 2021), DiaRNN (Majumder et al., 2019), COSMIC (Ghosal et al., 2020), Af-CAN (Wang et al., 2021), AGHMN (Jiao et al., 2020), COGMEN and RGAT (Ishiwatari et al., 2020) as our

[4] scipy.stats.shapiro.html

baselines. For MELD, we selected DiaGCN (Ghosal et al., 2019), DiaCRN (Hu et al., 2021), MMGCN (Wei et al., 2019), UniMSE (Hu et al., 2022b), COGMEN and MM-DFN (Hu et al., 2022a) as baselines. For MOSEI, we selected Mul-Net (Shenoy et al., 2020), TBJE (Delbrouck et al., 2020), COGMEN and MR (Tsai et al., 2020).

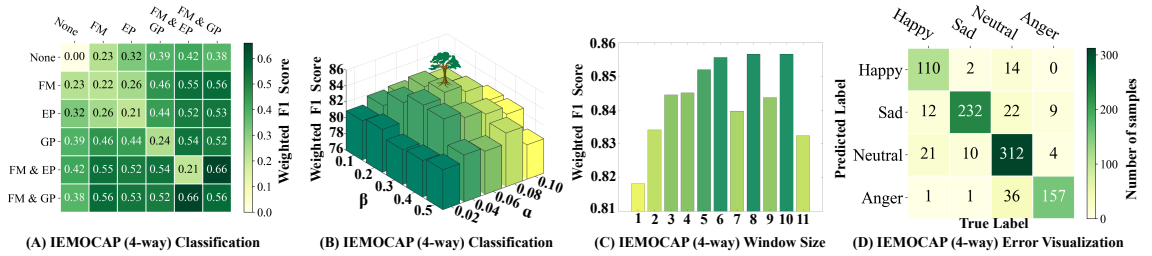


Figure 5.5: (A) WF1 gain with different augmentation pairs; (B∼C) Parameter tuning; (D) Imbalanced dataset.

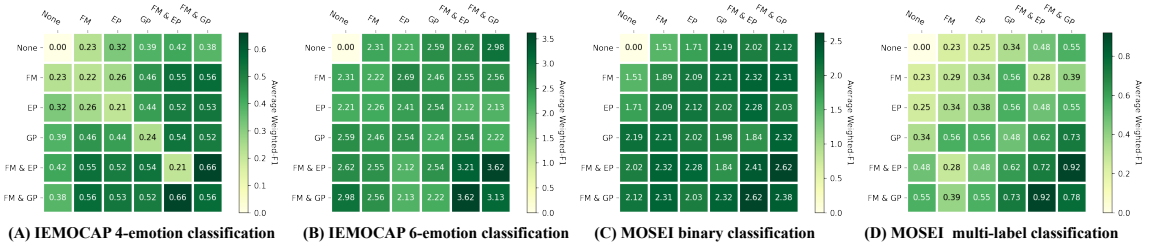


Figure 5.6: Average WF1 gain when contrasting different augmentation pairs, compared with training without graph augmentation module.

5.4.4 Parameter Sensitive Study

We first examined whether applying different data augmentation methods improves JOYFUL. We observed in Figure 5.5 (A) and Figure 5.6, when we consider the combinations of (FM & EP) and (FP & GP) as two graph augmentation methods of the original graph, we could achieve the best performance. Furthermore, we have the following observations:

(1) **Obs.1: Graph augmentations are crucial.** Without any data augmentation, the GCL module will not improve accuracy, judging from the averaged WF1 gain of the pair (None, None) in the upper left corners of Figure 5.6. In contrast, composing

an original graph and its appropriate augmentation can benefit the averaged WF1 of emotion recognition, judging from the pairs (none, any) in the top rows or the left-most columns of Figure 5.6. Similar observations were made in graphCL (You et al., 2020), without augmentation, GCL simply compares two original samples as a negative pair with the positive pair loss becoming zero, which leads to homogeneous pushes of all graph representations away from each other. Appropriate augmentations can enforce the model to learn representations invariant to the desired perturbations by maximizing the agreement between a graph and its augmentation.

(2) **Obs.2: Composing different augmentations benefits the model's performance more.** Applying augmentation pairs of the same type often does not result in the best performance (see diagonals in Figure 5.6). In contrast, applying augmentation pairs of different types results in a better performance gain (see off-diagonals of Figure 5.6). Similar observations were made in SimCSE (Gao et al., 2021). As mentioned in that study, the composition of augmentation pairs of different types corresponds to a "harder" contrastive prediction task, which could enable the learning of more generalizable representations.

(3) **Obs.3: One view having two augmentations results in better performance.** Generating each view by two augmentations further improves performance, ie., the augmentations FM & EP, or FM & GP. The augmentation pair (FM & EP, FM & GP) results in the largest performance gain compared to other augmentation pairs. We conjectured that the reason is that simultaneously changing structural and attribute information of the original graph can obtain more heterogeneous contextual information for nodes, which can be considered as "harder" example to prompt the GCL model to obtain more generalizable and robust representations.

Thus, we selected (FM & EP) and (FM & GP) as the default augmentation strategy since they achieved the best performance.

JOYFUL has three hyperparameters. α and β determine the importance of MF and GCL in Eq.(5.17), and window size controls the contextual length of conversations. Specifically, as shown in Figure 5.5 (B) and Figure 5.7, we observed how α and β affect the performance of JOYFUL by varying α from 0.02 to 0.10 in 0.02 intervals and β from 0.1 to 0.5 in 0.1 intervals.

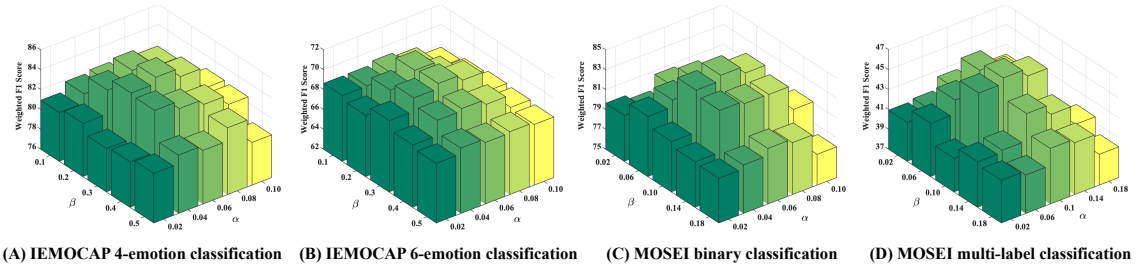


Figure 5.7: Parameters tuning for α and β on validation datasets for all multimodal emotion recognition tasks.

The results indicate that JOYFUL achieved the best performance when $\alpha \in [0.06, 0.08]$ and $\beta \in [0.2, 0.3]$ on IEMOCAP and when $\alpha \in [0.06, 0.1]$ and $\beta = 0.1$ on MOSEI. The reason why these parameters can affect the results is that when $\alpha < 0.06$, MF becomes weaker and representations contain too much noise, which cannot provide a good initialization for downstream MERC tasks. When $\alpha > 0.1$, it tends to make reconstruction loss more important and JOYFUL tends to extract more common features among multiple modalities and loses attention to explore features from unimodal. When β is small, graph contrastive loss becomes weaker, which leads to indistinguishable representation. A larger β wakes the effect of MF, leading to a local optimal solution. We set α=0.06 and β=0.3 for IEMOCAP and MELD. We set α=0.06 and β =0.1 for MOSEI.

Figure 5.5 (C), Tables 5.7 and 5.8 show that when the window size $\in [6, 8]$ for IEMOCAP (6-way) and the window size is 6 for IEMOCAP (4-way), JOYFUL achieved the best performance. A small window size will miss much contextual information, and a large-scale window size contains too much noise. We set the window size for the past and future to 6.

5.5 Experimental Results and Discussion

5.5.1 Performance of JOYFUL

Tables 5.9 and 5.10 show that JOYFUL outperformed all baselines in terms of accuracy and WF1, improving 5.0% and 1.3% in WF1 for 6-way and 4-way, respectively. Graph-based methods, COGMEN and JOYFUL, outperform Transformers-based methods, Mult and FE2E.

P&F	*Happiness*	*Sadness*	*Neutral*	*Anger*	*Accuracy*	*WF1*
size=1	83.27	83.04	80.63	81.54	81.87	81.82
size=2	79.02	82.92	83.93	86.65	83.46	83.41
size=3	80.88	86.34	84.07	85.64	84.52	84.45
size=4	**83.92**	85.83	83.91	84.35	84.52	84.51
size=5	82.93	87.85	83.79	86.47	85.26	85.20
size=6	81.73	86.42	85.17	88.46	85.58	85.56
size=7	79.33	86.07	83.29	86.40	83.99	83.97
size=8	80.14	**88.11**	85.06	88.15	**85.68**	**85.66**
size=9	77.29	87.85	83.56	87.19	84.41	84.37
size=10	80.00	87.47	**85.29**	**88.64**	85.68	85.66
size=ALL	79.87	84.35	83.20	84.75	83.24	83.24

Table 5.7: Results for various window sizes for graph formation on the IEMOCAP (4-way).

P&F	*Hap.*	*Sad.*	*Neu.*	*Ang.*	*Exc.*	*Fru.*	*Acc.*	*WF1*
size=1	57.85	80.43	62.88	60.61	70.76	60.99	65.50	65.85
size=2	56.27	79.57	64.17	60.87	72.50	61.52	65.93	66.36
size=3	60.80	80.26	66.06	64.47	73.17	62.70	67.71	68.09
size=4	59.95	80.79	67.96	67.18	71.60	64.89	68.64	69.05
size=5	60.06	81.42	68.23	66.33	73.88	63.24	68.76	69.17
size=6	**60.94**	**84.42**	68.24	**69.95**	73.54	**67.55**	**70.55**	**71.03**
size=7	59.84	80.53	67.93	68.12	73.72	63.91	68.82	69.26
size=8	57.66	82.17	**70.56**	67.53	73.92	64.79	69.75	70.12
size=9	58.01	81.13	70.22	65.42	**75.05**	61.49	68.82	69.12
size=10	59.77	81.84	69.17	65.85	73.56	63.51	68.95	69.38
size=ALL	54.74	78.75	66.58	64.56	68.63	63.46	66.42	66.80

Table 5.8: Results for various window sizes for graph formation on the IEMOCAP (6-way).

Transformers-based methods cannot distinguish intra- and inter-speaker dependencies, distracting their attention to important utterances. Furthermore, they use the cross-modal attention layer, which can enhance common features among modalities while losing unimodal specific features (Rajan et al., 2022). JOYFUL outperforms other GNN-based methods since it explored features from both the contextual and specific levels, and used GCL to obtain more distinguishable features. However, JOYFUL cannot improve in Happy for 4-way and in Excited for 6-way since samples in IEMOCAP were insufficient for distinguishing

these similar emotions (Happy is 1/3 of Neutral in Fig. 5.5 (D)). Without labels' guidance to re-sample or re-weight the underrepresented samples, self-supervised GCL, utilized in JOYFUL, cannot ensure distinguishable representations for samples of minor classes by only exploring graph topological information and vertex attributes.

Method	*IEMOCAP 6-way (F1)* ↑						*Average* ↑	
	Hap.	*Sad.*	*Neu.*	*Ang.*	*Exc.*	*Fru.*	*Acc.*	*WF1*
Mult	48.23	76.54	52.38	60.04	54.71	57.51	58.04	58.10
FE2E	44.82	64.98	56.09	62.12	61.02	57.14	58.30	57.69
DiaRNN	32.88	78.08	59.11	63.38	73.66	59.41	63.34	62.85
COSMIC	53.23	78.43	62.08	65.87	69.60	61.39	64.88	65.38
Af-CAN	37.01	72.13	60.72	67.34	66.51	66.13	64.62	63.74
AGHMN	52.10	73.30	58.40	61.91	69.72	62.31	63.58	63.54
RGAT	51.62	77.32	65.42	63.01	67.95	61.23	65.55	65.22
COGMEN	51.91	81.72	**68.61**	66.02	**75.31**	58.23	68.26	67.63
JOYFUL	**60.94**†	**84.42**†	68.24	**69.95**†	73.54	**67.55**†	**70.55**†	**71.03**†

Table 5.9: Overall performance comparison on IEMOCAP (6-way) in the multimodal (A+T+V) setting. Symbol † indicates that JOYFUL significantly surpassed all baselines using t-test with $p < 0.005$.

Method	*Happy*	*Sadness*	*Neutral*	*Anger*	*WF1*
Mult	88.4	86.3	70.5	87.3	80.4
RAVEN	86.2	83.2	69.4	86.5	78.6
MTAG	85.9	80.1	64.2	76.8	73.9
PMR	**89.2**	87.1	71.3	87.3	81.0
MICA	83.7	75.5	61.8	72.6	70.7
COGMEN	78.8	86.8	84.6	88.0	84.9
JOYFUL	80.1	**88.1**†	**85.1**†	**88.1**†	**85.7**†

Table 5.10: Overall performance comparison on IEMOCAP (4-way) in the multimodal (A+T+V) setting.

Tables 5.11 & 5.12 show that JOYFUL outperformed the baselines in more complex scenes with multiple speakers and various emotional labels. Compared with COGMEN and MM-DFN, which directly aggregate multimodal features, JOYFUL can fully explore

features from unimodal by specific representation learning to improve the performance. GCL module can aggregate similar emotional features for utterances to obtain better performance for multi-label classification. We cannot improve Happy in MOSEI since the samples are imbalanced and Happy has 1/6 number of Surprise, making JOYFUL difficult to identify.

Methods	**Emotion Categories of MELD (F1) ↑**					**Average ↑**	
	Neu.	***Sur.***	***Sad.***	***Joy***	***Anger***	***Acc.***	***WF1***
DiaGCN	75.97	46.05	19.60	51.20	40.83	58.62	56.36
DiaCRN	77.01	50.10	26.63	52.77	45.15	61.11	58.67
MMGCN	76.33	48.15	26.74	53.02	46.09	60.42	58.31
UniMSE	74.61	48.21	31.15	54.04	45.26	59.39	58.19
COGMEN	75.31	46.75	33.52	54.98	45.81	58.35	58.66
MM-DFN	**77.76**	50.69	22.93	54.78	47.82	62.49	59.46
JOYFUL	76.80	**51.91**†	**41.78**†	**56.89**†	**50.71**†	**62.53**†	**61.77**†

Table 5.11: Results on MELD with the multimodal setting. Underline indicates our reproduced results.

Method	***Happy***	***Sadness***	***Anger***	***Fear***	***Disgust***	***Surprise***
	Binary Classification (F1) ↑					
Mul-Net	67.9	65.5	67.2	87.6	74.7	86.0
TBJE	63.8	68.0	74.9	84.1	83.8	86.1
MR	65.9	66.7	71.0	85.9	80.4	85.9
COGMEN	70.4	72.3	76.2	88.1	83.7	85.3
JOYFUL	**71.7**†	**73.4**†	**78.9**†	**88.2**	**85.1**†	**86.1**
	Multi-label Classification (F1) ↑					
Mul-Net	70.8	70.9	74.5	86.2	83.6	87.7
TBJE	68.4	73.9	74.4	86.3	83.1	86.6
MR	69.6	72.2	72.8	86.5	82.5	87.9
COGMEN	**72.7**	73.9	78.0	86.7	85.5	88.3
JOYFUL	70.9	**74.6**†	**78.1**†	**89.4**†	**86.8**†	**90.5**†

Table 5.12: Results on MOSEI with the multimodal setting.

Modality	*IEMOCAP-4*		*IEMOCAP-6*		*MOSEI (WF1)*	
	Acc.	*WF1*	*Acc.*	*WF1*	*Binary*	*Multi-label*
Audio	64.8	63.3	49.2	48.0	51.2	53.3
Text	83.0	83.0	67.4	67.5	73.6	73.9
Video	44.6	43.4	28.2	28.6	23.6	24.4
A+T	82.6	82.5	67.5	67.8	74.7	74.9
A+V	68.0	67.5	52.7	52.5	61.7	62.4
T+V	80.0	80.0	65.2	65.5	73.1	73.4
w/o MF(B1)	85.3	85.4	70.0	70.3	76.2	76.5
w/o MF(B2)	85.2	85.1	69.2	69.5	75.8	76.2
w/o MF	85.2	84.9	69.0	69.2	75.4	75.8
COGMEN w/o GNN	80.1	80.2	62.7	62.9	72.3	72.9
w/o GCL	84.7	84.7	66.1	66.5	73.8	73.4
JOYFUL	**85.6**†	**85.7**†	**70.5**†	**71.0**†	**76.9**†	**77.2**†

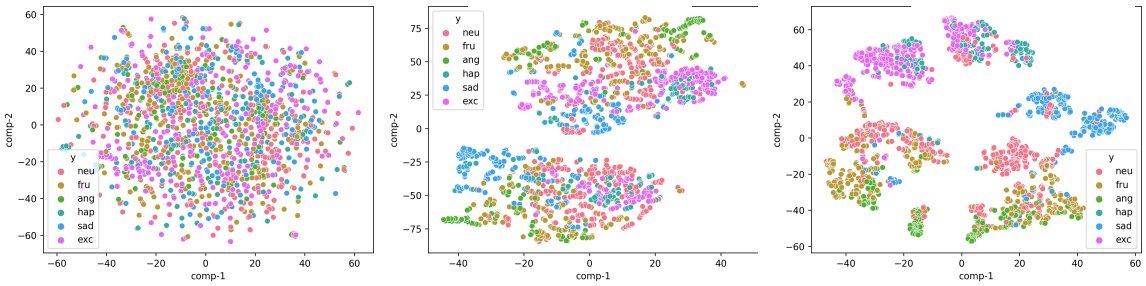
Table 5.13: Ablation study with different modalities.

5.5.2 Ablation Study

To verify the performance gain from each component, we conducted additional ablation studies. Table 5.13 shows multimodalities can greatly improve JOYFUL's performance compared with each single modality. GCL and each component of MF can separately improve the performance of JOYFUL, showing their effectiveness. JOYFUL w/o GCL and COGMEN w/o GNN utilize only a multimodal fusion mechanism for classification without additional modules to optimize node representations. The comparison between them demonstrates the effectiveness of the multimodal fusion mechanism in JOYFUL.

We visualized the node features to understand the function of the multimodal fusion mechanism and the GCL-based node representation learning component, as shown in Figure 5.8. Figure 5.8 (A) shows the concatenated multimodal features on the input side. Figure 5.8 (B) shows the representation of utterances after the feature fusion module. Figure 5.8 (C) shows the representation of the utterances after the GCL module (Eq.(10)) and before the pre-softmax layer (Eq.(11)). We observed that utterances could be roughly separated after the feature fusion mechanism, which indicates that the multimodal fusion mechanism can learn distinctive features to a certain extent. After GCL-based module,

JOYFUL can be easily separated, demonstrating that GCL can provide distinguishable representation by exploring vertex attributes, graph structure, and contextual information.



(A) Initial feature visualization **(B) Feature fusion visualization** **(C) Feature GCL visualization**

Figure 5.8: t-SNE visualization of IEMOCAP (6-way) features.

Method	***One-Layer (WF1)***		***Two-Layer (WF1)***		***Four-Layer (WF1)***	
	COGMEN	JOYFUL	COGMEN	JOYFUL	COGMEN	JOYFUL
Unattack	67.63	71.03	63.21	**71.05**	58.39	70.96
5% Noisy	65.26	**70.82**	61.35	70.55	56.28	70.10
10% Noisy	62.26	70.33	59.24	**70.45**	53.21	69.23
15% Noisy	57.28	**69.98**	55.18	69.21	52.32	67.96
20% Noisy	54.22	68.52	51.79	**68.82**	50.72	67.23

Table 5.14: Adversarial attacks for GNN with different depth on 6-way IEMOCAP.

5.5.3 Over-Smoothing

We deepened the GNN layers to verify JOYFUL's ability to alleviate the over-smoothing. In Table 5.14, COGMEN with four-layer GNN was 9.24% lower than that with one layer, demonstrating that over-smoothing decreases performance, while JOYFUL alleviated this issue. To verify robustness, following Tan et al. (2022), we randomly added 5%∼20% noisy edges to the training data. In Table 5.14, COGMEN was easily affected by the noise, decreasing 10.8% performance in average with 20% noisy edges, while JOYFUL had strong robustness with only an average 2.8% performance reduction for 20% noisy edges.

Method	*Modality*	*WF1*
	IEMOCAP 6-way	
CESTa	Text	67.10
SumAggGIN	Text	66.61
DiaCRN	Text	66.20
DialogXL	Text	65.94
DiaGCN	Text	64.18
COGMEN	Text	66.00
DAG-ERC	Fine-tune Text (RoBERTa-large)	68.03
JOYFUL	Text (Sentence-BERT)	67.48
	Text (RoBERTa-large)	68.05
	Fine-tune Text (RoBERTa-large)	**68.45**
	A+T+V	**71.03**

Table 5.15: Overall performance comparison on MOSEI with Text Modality.

5.5.4 Unimodal Performance

The focus of this study was multimodal emotion recognition. However, we also compared JOYFUL with unimodal methods to evaluate its performance of JOYFUL. We compared it with DAG-ERC (Shen et al., 2021b), CESTa (Wang et al., 2020c), SumAggGIN (Sheng et al., 2020), DiaCRN (Hu et al., 2021), DialogXL (Shen et al., 2021a), DiaGCN (Ghosal et al., 2019), and COGMEN (Joshi et al., 2022). Following COGMEN, text-based models were specifically optimized for text modalities and incorporated changes to architectures to cater to text. As shown in Table 5.15, JOYFUL, being a fairly generic architecture, still achieved better or comparable performance with respect to the state-of-the-art unimodal methods. Adding more information via other modalities helped to further improve the performance of JOYFUL (Text vs A+T+V). When using only text modality, the DAG-ERC baseline could achieve higher WF1 than JOYFUL. And we conjecture the main reasons is: DAG-ERC (Shen et al., 2021b) fine-tuned RoBERTa large model (Liu et al., 2019), with 354 million parameters, as their text encoder. By fine-tuning on RoBERTa large model under the guidance of downstream emotion recognition signals, RoBERTa large model can provide the most suitable text features for ERC. Compared with DAG-ERC, JOYFUL and other methods directly use Sentence-BERT (Reimers and Gurevych, 2019), with 110 million parameters,

as their text encoder without fine-tuning on ERC datasets.

To verify whether the above inference is reasonable, we used RoBERTa large model as our text feature extractor called *Text (RoBERTa-large)*. And we fine-tuned RoBERTa large model on the downstream IEMPCAP (6-way) dataset, following the same method of DAG-ERC called *Fine-tune Text (RoBERTA-large)*. The observation meets our intuition. With RoBERTa large model, JOYFUL improved the performance (68.05 vs 67.48) compared with Sentence-BERT as our text encoder. And JOYFUL could obtain better performance (68.45 vs 68.03) in terms of WF1 than DAG-ERC with fine-tuned RoBERTa-large, demonstrating that fine-tuning large-scale model can help obtain richer text features to improve the performance. However, considering a fair comparison with other multimodal emotion recognition baselines (they do not have the fine-tuning process (Joshi et al., 2022; Ghosal et al., 2019)) and saving the additional time-consuming on fine-tuning, we directly adopt Sentence-BERT as our text encoder for IEMOCAP.

5.5.5 Case Study

To show the distinguishability of the node representations, we visualize the node representations of FE2E, COGMEN, and JOYFUL on 6-way IEMOCAP. In Figure 5.9, COGMEN and JOYFUL obtained more distinguishable node representations than FE2E, demonstrating that graph structure is more suitable for MERC than Transformers. JOYFUL performed better than COGMEN, illustrating the effectiveness of GCL. In Figure 5.10, we randomly sampled one example from each emotion of IEMOCAP (6-way) and chose best-performing COGMEN for comparison. JOYFUL obtained more discriminate prediction scores among emotion classes, showing GCL can push samples from different emotion class farther apart.

5.5.6 Multimodal Sentiment Analysis

We conducted experiments on two publicly available datasets, **MOSI** (Zadeh et al., 2016b) and **MOSEI** (Zadeh et al., 2018b), to investigate the performance of JOYFUL on the multimodal sentiment analysis (MSA) task.

Datasets: MOSI contains 2,199 utterance video segments, and each segment is manually annotated with a sentiment score ranging from -3 to +3 to indicate the sentiment polarity

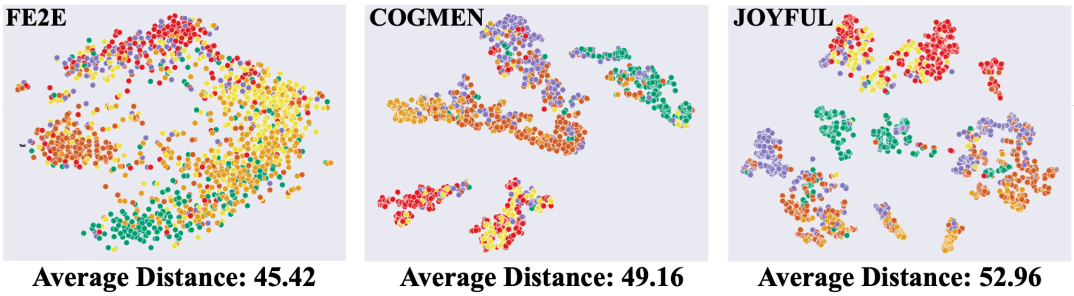


Figure 5.9: t-SNE visualization of IEMOCAP (6-way).

and relative sentiment strength of the segment. MOSEI contains 22,856 movie review clips from the YouTube website. Each clip is annotated with a sentiment score and an emotion label. And the exact number of samples for training/validation/test are 1,284/229/686 for MOSI and 16,326/1,871/4,659 for MOSEI.

Metrics: Following previous studies (Han et al., 2021a; Yu et al., 2021), we utilized evaluation metrics: mean absolute error (MAE) measures the absolute error between predicted and true values. Person correlation (Corr) measures the degree of prediction skew. Seven-class classification accuracy (ACC-7) indicates the proportion of predictions that correctly fall into the same interval of seven intervals between -3 and +3 as the corresponding truths. And binary classification accuracy (ACC-2) was computed for non-negative/negative classification results.

Baselines: We compared JOYFUL with three types of advanced multimodal fusion frameworks for the MSA task as follows, including current SOTA baselines MMIM (Han et al., 2021b) and BBFN (Han et al., 2021a): (1) Early multimodal fusion methods, which combine the different modalities before they are processed by any neural network models. We utilized Multimodal Factorization Model (**MFM**) (Tsai et al., 2019b), and Multimodal Adaptation Gate BERT (**MAG-BERT**) (Rahman et al., 2020) as baselines. (2) Late multimodal fusion methods, which combine the different modalities before the final decision or prediction layer. We utilized multimodal Transformer (**MuIT**) (Tsai et al., 2019a), and modal-temporal attention graph (**MTAG**) (Yang et al., 2021) as baselines. (3) Hybrid multimodal fusion methods combine early and late multimodal fusion mechanisms to capture the consistency

Figure 5.10: Visualization of emotion probability, each first row is JOYFUL and each second row is COGMEN.

and the difference between different modalities. We utilized modality-invariant and modality-specific representations for MSA (**MISA**) (Hazarika et al., 2020), Self-Supervised multi-task learning for MSA (**Self-MM**) (Yu et al., 2021), Bi-Bimodal Fusion Network (**BBFN**) (Han et al., 2021a), and MultiModal InfoMax (**MMIM**) (Han et al., 2021b) as baselines.

Implementation Details: The results of proposed JOYFUL were averaged over ten runs using random seeds. We keep all hyper-parameters and implementations the same as in the MERC task. To make JOYFUL fit the MSA task, we replace the current cross-entropy loss $\mathcal{L}_{ce}$ in Eq. (15) by mean absolute error loss $\mathcal{L}_{mae}$ as follows:

$$\mathcal{L}_{mae} = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}_i - y_i|, \tag{5.18}$$

where $\hat{y}_i$ is the predicted value for the i-th sample, y_i is the truth label for the i-th label, m is the total number of samples, and $|\cdot|$ is the L_1 norm. We denote this model as JOYFUL+MAE.

Experimental results on the MOSI and MOSEI datasets are listed in Table 5.16. Although the proposed JOYFUL could outperform most of the baselines (above the blue line),

Method	MOSI				MOSEI			
	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 ↑	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 ↑
MFM	0.877	0.706	35.4	81.7	0.568	0.717	51.3	84.4
MAG-BERT	0.731	0.789	✗	84.3	0.539	0.753	✗	85.2
MulT	0.861	0.711	✗	84.1	0.580	0.703	✗	82.5
MTAG	0.866	0.722	0.389	82.3	✗	✗	✗	✗
MISA	0.804	0.764	✗	82.10	0.568	0.724	✗	84.2
Self-MM	0.713	0.789	✗	85.98	0.530	0.765	✗	85.17
BBFN	0.776	0.755	45.00	84.30	0.529	0.767	**54.80**	**86.20**
MMIM	**0.700**	**0.800**	**46.65**	**86.06**	**0.526**	**0.772**	54.24	85.97
JOYFUL+MAE	0.711	0.792	45.58	85.87	0.529	0.768	53.94	85.68

Table 5.16: Experimental results on the MOSI and MOSEI datasets. ✗ indicates unreported results. **Bold** indicates the least MAE, highest other scores for each dataset.

Case	**Input modality**			**Target**	
	Text	Visual	Acoustic	MSA	MERC
Case A	Plot to it than that the action scenes were my favorite parts through it's.	Smiling face Relaxed wink	Stress Pitch variation	+1.666	Positive
Case B	You must promise me that you'll survive, you won't give up.	Full of tears in his eyes	The voice is weak and trembling	-1.200	Negative

Table 5.17: Case study on the importance of each modality for MSA and MERC tasks. Blue in Text modality marks the contents including the strength of sentiments. Underline marks fragments contributing to the target on MERC.

it performs worse than current SOTA models: BBFN and MMIM (below the blue line). We conjecture the main reasons are: when determining the strength of sentiments, compared with visual and acoustic modalities that may contain much noise data, text modality is more important for prediction (Han et al., 2021a). Table 5.17 lists such examples, where textual modality is more indicative than other modalities for the MSA task. Because the two baselines: BBFN (Han et al., 2021a) and MMIN (Han et al., 2021b), pay more attention to the text modality than visual and acoustic modalities during multimodal feature fusion, they

may achieve low MAE, high Corr, Acc-2, and Acc-7. Specifically, BBFN (Han et al., 2021a) proposed a Bi-bimodal fusion network to enhance the text modality's importance by only considered text-visual and text-acoustic interaction for features fusion. Conversely, considering the three modalities are all important for the MERC task as presented in Table 5.17, we designed JOYFUL to utilize the concatenation of the three modalities representations for prediction. Similar to our proposal, MISA and MAG-BERT considered the three modalities equally important during feature fusion but performed worse than SOTA baselines on the MSA task. In our analysis, due to this attention to the modalities, JOYFUL outperformed the SOTA baselines in the MERC but underperformed SOTA baselines on the MSA.

5.6 Limitations and Summary

5.6.1 Limitations

JOYFUL has a limited ability to classify minority classes with fewer samples in unbalanced datasets. Although we utilized self-supervised graph contrastive learning to learn a distinguishable representation for each utterance by exploring vertex attributes, graph structure, and contextual information, GCL failed to separate classes with fewer samples from the ones with more samples because the utilized self-supervised learning lacks the label information and does not balance the label distribution. Another limitation of JOYFUL is that its framework was designed specifically for multimodal emotion recognition tasks, which is not straightforward and general as language models (Devlin et al., 2019; Liu et al., 2019) or image processing techniques (LeCun and Bengio, 1995). This setting may limit the applications of JOYFUL for other multimodal tasks, such as the multimodal sentiment analysis task and the multimodal retrieval task. Finally, although JOYFUL achieved SOTA performances on three widely-used MERC benchmark datasets, its performance on larger-scale and more heterogeneous data in real-world scenarios is still unclear.

5.6.2 Summary

We proposed a joint learning model (JOYFUL) for MERC, that involves a new multimodal fusion mechanism and GCL module to effectively improve the performance of MERC. The

MR mechanism can extract and fuse contextual and unimodal specific emotion features, and the GCL module can help learn more distinguishable representations. For future work, we plan to investigate the performance of using supervised GCL for JOYFUL on unbalanced and small-scale emotional datasets.

Chapter 6

The Future of Emotion Recognition

In the previous chapter, we described how neural networks have succeeded in current emotion recognition benchmarks and highlighted their key insights. Despite rapid progress, there is still a long way to go toward achieving genuine human-level emotion recognition. In this chapter, we discuss future work and open questions.

First, we discuss the main issues with current datasets in Section 6.1. Specifically, we examine the scarcity of training data in Section 6.1.1, annotation and diversity of datasets in Section 6.1.2, and the presence of noisy data in Section 6.1.3.

Then, we introduce the challenges and future work related to emotion recognition models in Section 6.2. We discuss the generalization ability of models in Section 6.2.1, multimodal fusion in Section 6.2.2, unbiased emotion learning in Section 6.2.3, and incomplete multimodal conversation emotion recognition in Section 6.2.4.

Finally, we review several important research questions in this field in Section 6.3. Specifically, we explore the efficiency of emotion recognition in complex real-world scenarios in Section 6.3.1, zero-shot multimodal conversation emotion recognition in Section 6.3.2, multi-label emotion reasoning in Section 6.3.3 and human-robot interaction in Section 6.3.4.

6.1 Future Work: Datasets

In this section, we list some important issues in current multimodal emotion recognition datasets. Solving all of these issues is one of the research motivations for our future research.

6.1.1 Scarcity of Training Data

To achieve greater accuracy and better outcomes, deep learning models typically require abundant data and substantial computational resources for training. However, obtaining extensive annotated data and conducting model training for multimodal emotion recognition is a challenging and costly endeavor. Multimodal conversational emotion recognition models need sufficient and comprehensive emotional samples to accurately predict or classify emotions. The existing multimodal benchmark datasets, such as IEMOCAP, MELD and SEMAINE, contain only 11,098, 5,810, and 394 utterances, respectively. Although it is possible to collect large amounts of multimodal conversation data from sources such as social media, the process of emotion labeling is often expensive and time consuming. Furthermore, collected data often suffer from issues such as ambiguous or multiple labels, making it difficult to obtain sufficient labeled data and leading to a scarcity of training data. This scarcity limits the effectiveness of current multimodal conversational emotion recognition models. In addition, ample computational resources and storage space are essential for large-scale training. Consequently, effectively utilizing limited data and computational resources, as well as accelerating the training process, remains a significant challenge.

6.1.2 Annotation and Diversity of datasets

Datasets are crucial for the performance and generalization ability of deep learning models. An ideal dataset should be representative, diverse, and of sufficient scale while maintaining high-quality labels. Datasets enable models to learn patterns through sample observation, and diverse data provide learning opportunities in different contexts. Large-scale datasets can mitigate overfitting issues, and high-quality labels offer accurate supervision signals. Therefore, constructing a large-scale, diverse dataset is imperative in the MER domain. However, annotating multimodal data requires professionals to subjectively evaluate text, speech, and images, which is both time-consuming and expensive. The challenge lies in developing a high-quality, large-scale, and diverse MER dataset.

6.1.3 Noisy and Unbalanced Dataset

Multimodal conversation emotion recognition models need to effectively eliminate heterogeneity and noise between modalities to achieve accurate emotion prediction or classification. Multimodal data are naturally heterogeneous, with significant differences in processing methods and representation forms between modalities. Additionally, multimodal conversation data often contain a large amount of redundant or noisy information, while emotion is typically determined by a small amount of consistent key information, such as specific words in a sentence, a particular frequency band in speech, or a distinct expression in a video. In some extreme cases, part of the modal information may be rendered unusable due to noise interference, such as ambiguous sentence expressions, noisy speech, or obstructed facial expressions. Therefore, the heterogeneity and noise of the data limit the effectiveness of current multimodal conversational emotion recognition models.

Moreover, multimodal dialogue data samples often suffer from serious imbalance issues, which interfere with the unbiased learning of models. These models rely on cross-modal feature fusion and are driven by emotion category sample data, making them sensitive to the number of emotion category samples. However, multimodal conversation emotion data naturally exhibit category sample imbalance, where a few emotion categories dominate, and most categories are underrepresented. For example, in the MELD dataset, the "fear" emotion accounts for only 1.91% of the total samples, and "disgust" accounts for just 2.61%. A similar sample distribution is observed in the SEMAINE benchmark dataset. The scarcity of samples from certain emotion categories makes it difficult for the model to learn unbiasedly, significantly affecting its prediction accuracy for these underrepresented emotional categories. Consequently, the unbalanced sample distribution limits the effectiveness of current multimodal conversational emotion recognition models.

6.2 Future Work: Models

6.2.1 Generalization Ability of Models

Despite the proposal of numerous excellent MER models, they are usually trained on specific datasets that rely on nonrealistic scenarios, making them difficult to adapt to industrial

applications. Therefore, MER models need to possess strong generalization capabilities to be applicable to different scenarios and tasks. However, due to limitations in datasets and the complexity of models, the generalization capability of these models with regard to new domains or unseen data remains a challenge. Constructing more universal and transferable models is a significant challenge that must be tackled.

6.2.2 Multimodal Fusion

Multimodal feature fusion is crucial for multimodal conversational emotion recognition task. The fused feature vector can represent the consistent semantics and complementary information between modalities. However, there are many different information interactions between the modalities, with numerous consistent or complementary features hidden in multiple time series or local spatial correlations. Since multimodal conversation data are heterogeneous and contain noise, there are significant differences in the temporal periods and spatial distributions of different modal features, and the spatio-temporal importance between modalities is dynamic. Few works currently address these differences, and more efforts are needed for the deep fusion of multimodal features. The combination of data from disparate modalities for emotion recognition is a challenging task. Temporal misalignment and heterogeneities between features in different modalities can complicate the fusion process, often preventing models from fully utilizing the additional information of the various modalities. By quantifying the information content in various modalities through entropy, it becomes feasible to evaluate the uncertainty or predictability of each data source. This approach can pinpoint modalities that provide substantial information, while also identifying those that introduce noise or redundancy. Furthermore, employing concepts such as mutual information can elucidate the extent of interdependence between modalities. These insights facilitate a more harmonious and informed fusion process, ensuring that the relationships and synergies between the modalities are optimally leveraged. Therefore, harnessing information theory to develop a powerful strategy for amalgamating textual, facial, and auditory features and seamlessly incorporating this multimodal information into a comprehensive model is of paramount importance.

6.2.3 Unbiased Emotional Learning

Many benchmark datasets in the field of multimodal conversational emotion recognition suffer from serious sample category imbalance. In these datasets, minority emotion categories contain a large amount of data, while majority emotion categories have only a small amount. In the case of unbalanced data, existing models tend to be biased towards fitting the minority emotions with large amounts of data, leading to insufficient learning on the majority emotions with small data samples. This imbalance results in poor recognition accuracy for most emotion categories. Therefore, addressing the small sample problem in multimodal dialogue emotion recognition urgently requires further research.

6.2.4 Incomplete Multimodal Conversation Emotion Recognition

In real-world scenarios, each modality is not always available, leading to modal incompleteness problems. For example, noise can interfere with voice data, expressions can be blocked, or lighting can be dim. In such cases, some modal information becomes unavailable due to noise interference. This need for modal integrity reduces the applicability of multimodal conversation emotion recognition methods. Therefore, developing cross-modal content recovery methods based on deep learning is essential to achieve effective emotion recognition even when some modalities are missing.

6.3 Research Questions

In the last section, we discuss some central research questions in this field that still remain open and yet to be answered in the future.

6.3.1 Efficiency in Complex Real-World Scenes

6.3.1.1 Extension of AFTER to Multiple Annotators

Following previous works (Xu et al., 2013; Zhang et al., 2022d), we assumed that each sample is annotated with its ground truth labels. Thus, we first masked the labels on all training datasets and considered them as unlabeled data. Then, we unmasked the labels of

some samples for training if these samples were selected by active learning. However, in the real world, annotators with different knowledge, ages, genders, intuitions, backgrounds, and cultures (Bhardwaj et al., 2010; Dang et al., 2010) may annotate the same sample differently.

Following previous studies, such as learning from the soft label (Peterson et al., 2019; Uma et al., 2020; Fornaciari et al., 2021) and learning from the hard label of individual annotators (Cohn and Specia, 2013; Rodrigues and Pereira, 2018; Davani et al., 2022), we extended our proposed method, AFTER, to address the aforementioned real-world situation by suggesting the following potential solutions:

1. Individual-level Entropy (indi): We can measure the reliability of each annotator by calculating the individual-level entropy for each annotator. Given the prediction label for sample x as $\mathbf{z}^a = [z_1^a, \cdots, z_n^a]$ by annotator a, the entropy can be calculated by

$$H_{indi}(p^a|x) = -\sum_{i=1}^{n} p_i^a(x) \log(p_i^a(x)), \tag{6.1}$$

where $p_i^a(x) = \text{softmax}(z_i^a(x))$. We select the $(instance, annotator)$ pair with the highest entropy using:

$$argmax_{x \in U, a \in A} H_{indi}(p^a|x), \tag{6.2}$$

where U denotes the unlabeled set and A denotes the annotator pool.

2. Group-level Entropy (group): Instead of focusing solely on individual uncertainty, we can query the data by considering the group-level uncertainty. To represent the uncertainty of the group on a sample, we calculate the entropy baseline based on the aggregation of each annotator's specific output. Therefore, we normalize and sum the logits of each annotator at the group level: $\mathbf{z}_{gropu} = [z_1, \cdots z_n] = \sum_{a=1}^{|A|} \mathbf{z}_{norm}^a$, and calculate the group-level entropy as follows:

$$H_{group}(x) = -\sum_{i=1}^{n} p_i(x) \log(p_i(x)), \tag{6.3}$$

where $p_i(x) = \text{softmax}(z_i(x))$ and $|A|$ represent the number of annotators. We then query the data with the highest group-level uncertainty.

3. Vote Variance (vote): Another method to measure the uncertainty among a group is by computing the variance of the votes. Given the prediction y^a of annotator a, we calculate the vote variance as follows:

$$\text{Var}(x) = \frac{1}{|A|}\sum_{a=1}^{|A|}(y^a - \mu)^2, \tag{6.4}$$

where $\mu = \frac{1}{|A|}\sum_{a=1}^{|A|} y^a$ and $|A|$ represents the number of annotators.

4. Mixture of Group and individual Entropy (mix): We also consider a variant that combines the group-level and individual-level entropy by simply adding the two $H_{mix} = H_{indi} + H_{group}$.

6.3.1.2 Extension of AFTER with Soft-labels

As the number of emotions increases, the difference in the results depending on the annotator becomes more pronounced. Generally, soft labels are used to address this issue. We conducted additional experiments on the IEMOCAP dataset, where each utterance was labeled by three human annotators. Furthermore, each annotator was allowed to choose more than one categorical label if they felt it necessary (Busso et al., 2008).

To simulate three different annotators, following (Fayek et al., 2016), we trained three separate DNNs on the IEMOCAP dataset using the hard labels from each annotator. Specifically, each DNN architecture contained seven feed-forward fully-connected layers and adopted ReLU as the activation function. The input layer's dimensionality was 2,624 (64 frames $\times$ 41 coefficients per frame) and the output layer is a four-way softmax layer, which produced the posterior class probabilities. We used the cross-entropy loss for the emotion recognition of each classifier:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} y_i^j \log(\hat{y}_i^j), \tag{6.5}$$

where c is the number of emotion classes, N denotes the total number of samples, $\hat{y}_i^j$ stands for the i-th predicted label, and y_i^j represents the i-th ground-truth label for the j-th class.

For more detailed implementations, please refer to (Fayek et al., 2016).

Then we obtained the soft labels for each speech sample by:

$$s = \frac{\sum_{a=1}^{A} h^{(n)}}{\sum_{i=1}^{c}\sum_{a=1}^{A} h^{(n)}}, \tag{6.6}$$

where s is a c-dimensional vector of soft labels, $h^{(n)}$ is a c-dimensional vector of one-of-c hard labels encoded from the n-th annotator, and A is the number of annotators. Table 6.1 illustrates several annotation examples from the IEMOCAP database labeled by three annotators and their corresponding labels to alleviate ambiguity.

Table 6.1: Examples of soft labels for IEMOCAP with three annotators. Annotation are in the form of (Annotator 1, Annotator 2, Annotator 3). Hard/Soft labels are in the form of [Anger, Happiness, Neutral, Sadness].

Annotation	Hard Label	Soft Label
(Anger, Anger, Anger)	[1,0,0,0]	[1,0,0,0]
(Happiness, Neutral, Neutral)	[0,0,1,0]	[0,0.33,0.66,0]
(Sadness, Sadness, Sadness;Neutral)	[0,0,0,1]	[0,0,0.25,0.75]

After obtaining the soft labels from the IEMOCAP datasets, AFTER also measures the uncertainty of each sample $\mathbf{x}_i$ as follows:

$$\text{Entropy}(\mathbf{x}_i) = -\sum_{j=1}^{c} P(\hat{y}_j|\mathbf{x}_i)\log P(\hat{y}_j|\mathbf{x}_i), \tag{6.7}$$

where c is the number of emotional classes and $P(\hat{y}_j|\mathbf{x}_i)$ represents the predicted probability of $\mathbf{x}_i$ for the j-th emotion. Following (Fayek et al., 2016), the classifier outputs the class with the highest posterior probability during evaluation. Experimental results in Table 6.2 demonstrate that AFTER outperformed all baselines even with soft labels, indicating its capability to handle real-world scenarios with complex soft-labeled emotions.

Table 6.2: Overall performance comparison on four emotion categories. AFTER adopted Entropy+Clustering and selected 20% samples for fine-tuning. Baselines use all samples from each corresponding dataset for training. THe symbol † indicates that AFTER significantly surpassed all baselines with $p < 0.05$ according to the t-test.

Methods	**IEMOCAP (hard-label)**		**IEMOCAP (soft-label)**	
	UA ↑	WA ↑	UA ↑	WA ↑
GLAM [2022]	74.01	72.98	68.15	64.33
LSSED [2021]	73.09	68.35	62.23	61.38
RH-emo [2022]	68.26	67.35	61.35	59.17
Pseudo-TAPT [2022]	74.30	70.26	69.98	68.23
w2v2-L-r-12 [2023]	74.28	70.23	70.31	69.24
AFTER	**76.07**†	**73.24**†	**73.37**†	**72.96**†

6.3.2 Zero-shot Multimodal Conversation Emotion Recognition

Due to the complexity of emotions and the high cost of labeling, it is difficult to fully annotate some emotional samples. Additionally, with the rapidly growing range of personal emotion annotations, real-world emotion recognition systems may frequently encounter unseen emotion labels. Therefore, improving the generalization performance of emotion recognition models is crucial. Deep learning methods utilizing zero-shot learning are expected to achieve better results in multimodal dialogue emotion recognition.

6.3.3 Multi-label emotion reasoning:

Multimodal conversation emotion recognition models need to establish accurate and consistent semantic associations and capture complementary semantic features between modalities. This approach can enhance emotional representation capabilities beyond what a single modality can achieve. However, unlike consistent semantics, complementary semantics highlight differences between modalities, which may include noise components. Therefore, balancing the relationship between consistency and semantics is a critical consideration at the model level. Constructing data samples requires agreement on exact labels from at least two different subjects. However, it is possible for people to experience multiple emotions simultaneously, such as feeling both sad and angry. Therefore, utilizing MemoR

for multi-label emotion reasoning holds promise as a future direction. In addition, future research areas include emotion detection based on physiological signals, audiovisual group emotion recognition, and driver gaze prediction, all of which are crucial for online learning and engagement prediction in real-world settings.

6.3.4 Human robot interaction (HRI)

Multimodal techniques will play a critical role in improving emotion identification performance compared to single-modal techniques. This necessitates the development of machine learning (ML) methods and deep learning (DL) architectures capable of handling heterogeneous data. Special attention must be paid to the data used for training and testing emotion recognition systems. Human-Robot Interaction (HRI) presents unique characteristics that can make data collected under controlled conditions or in diverse contexts unsuitable for real-world HRI applications. Therefore, there is a need to develop datasets specifically tailored for actual HRI scenarios in the future.

Chapter 7

Conclusions

In this dissertation, we gave readers a thorough overview of emotion recognition: the unimodal emotion recognition (PART I) and multimodal emotion recognition (PART II), as well as how we contributed to the development of this field.

In Chapter 2, we walked through the history of unimodal emotion recognition, which dates back to the 1970s. At the time, researchers already recognized its importance as a proper way of understanding human emotion and interaction. However, it was not until the 2015s that emotion recognition started to be formulated as a supervised learning problem by collecting large-scale human-labeled multiple unimodal training examples in the form. Since 2015, with the development of deep neural networks, the field has been moving strikingly fast. Innovations in building better datasets and more effective models have occurred alternately and both contributed to the development of the field. We also formally defined the task of emotion recognition and described the most commonly used datasets and evaluation metrics.

In Chapter 3, we proposed an active learning-based fine-tuning framework for speech emotion recognition, referred to as AFTER, which can be easily applied to noisy and heterogeneous real-world scenarios. Specifically, we first proposed an unsupervised task adaptation pre-training strategy to reduce the information gap between the pre-trained and downstream speech emotion recognition tasks, enabling the pre-trained model to understand the semantic information of the speech emotion recognition task. Then, we created two large-scale heterogeneous and noisy datasets to simulate real-world scenes. Furthermore,

we proposed AL strategies with clustering-based initialization to iteratively select a smaller, more informative, and diverse subset of samples for fine-tuning. This approach can efficiently eliminate noise and outliers, improve generalization, and reduce time consumption.

In PART II, we introduced multimodal emotion recognition methods and proposed joint multimodal fusion and graph contrastive learning for multimodal emotion recognition. Finally, we introduced the challenges and future work of multimodal emotion recognition.

In Chapter 4, we gave an overview of multimodal emotion recognition methods. Specifically, we first introduced the advantages and disadvantages of current multimodal fusion strategies: early fusion, late fusion, and hybrid fusion strategies. Next, we introduced the main backbones of multimodal emotion recognition, including deep neural networks, Seq2seq models, Transformers, and graph neural networks. We also summarized the widely used multimodal emotion recognition datasets and evaluation metrics.

In Chapter 5, we designed a new multimodal fusion mechanism that can simultaneously learn and fuse a global contextual representation and unimodal specific representations. We proposed a GCL-based cross-view framework to alleviate the difficulty of categorizing similar emotions, which helps to learn more distinctive representations of utterance by making samples with the same sentiment cohesive and those with different sentiments mutually exclusive. Extensive experiments conducted on three multimodal benchmark datasets demonstrated the effectiveness and robustness of our method.

In Chapter 6, we discussed future work and open questions in this field. We concluded the limitations from two aspects: datasets and models. Specifically, we summarized the limitations of current studies as: scarcity of large-scale training data, annotation and diversity of datasets, and noisy and unbalanced datasets. Furthermore, we summarize that current models have low generalization, low effectiveness in the multimodal fusion process, unbiased emotion learning scenes, and incomplete multimodal conversation emotion recognition.

All together, we are really excited about the progress that has been made in this field for the past 3 years and have been glad to be able to contribute to this field. At the same time, we also deeply believe that there is still a long way to go towards genuine human-level emotion recognition, and we are still facing enormous challenges and a lot of open questions that we will need to address in the future. One key challenge is that we still do not have good ways to approach real-world heterogeneous and noisy scenes. Although active learning is a

good method for identifying redundant and noisy data in real-world applications, how to effectively annotate samples and elegantly assign multiple labels to each utterance remains unclear. Multimodality can achieve better performance, while how to improve the efficiency of multimodal emotion recognition is still an open issue.

We also hope to encourage more researchers to work on the applications, or to apply emotion recognition to new domains or tasks. We believe that it will lead us towards building better human-machine interaction systems and hope to see these ideas implemented and developed in industry applications.

Bibliography

A.Adaeze, T.Noé, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *CoRR, abs/1806.09514*.

Arya Aftab, Alireza Morsali, Shahrokh Ghaemmaghami, and Benoit Champagne. 2022. A lightweight fully convolutional neural network for speech emotion recognition. In *Proceedings of ICASSP*, pages 6912–6916.

Parag Agrawal and Anshuman Suri. 2019. NELEC at semeval-2019 task 3: Think twice before going deep. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*, pages 266–271.

Md. Shad Akhtar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2018. A multi-task ensemble framework for emotion, sentiment and intensity prediction. *CoRR*, abs/1808.01216.

A.Kirsch, Jv.Amersfoort, and Y.Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Proceedings of NeurIPS*, pages 7024–7035.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *Proceedings of ICLR*.

Moataz M. H. El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.*, 44(3):572–587.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *CoRR, abs/2111.09296*.

Gilbert Badaro, Hussein Jundi, Hazem Hajj, and Wassim El-Hajj. 2018. EmoWordNet: Automatic expansion of emotion lexicon using English WordNet. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 86–93.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. *CoRR, abs/2202.03555*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of NeurIPS*, pages 12449–12460.

Guirong Bai, Shizhu He, Kang Liu, Jun Zhao, and Zaiqing Nie. 2020. Pre-trained language model based active learning for sentence matching. In *Proceedings of COLING*, pages 1495–1504.

Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. 2012. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision support systems*, 53(4):742–753.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of FG*, pages 59–66.

Pablo V. A. Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. 2018. The omg-emotion behavior dataset. In *Proceedings of IJCNN*, pages 1–7.

Murchana Baruah and Bonny Banerjee. 2022. Speech emotion recognition via generation using an attention-based variational recurrent neural network. In *Proceedings of INTERSPEECH*, pages 4710–4714.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society*, 57(1):289–300.

Vikas Bhardwaj, Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of LAW*, pages 47–55.

BJ.Abbaschian, D.Sierra-Sosa, and A.Elmaghraby. 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249–1258.

Smiley Blanton. 1915. The voice and the emotions. *Quarterly Journal of Speech*, 1(2):154–172.

Jonathan Boigne, Biman Liyanage, and Ted Östrem. 2020. Recognizing more emotions with less data using self-supervised transfer learning. *CoRR, abs/2011.05585*.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. 2005. A database of german emotional speech. In *Proceedings of INTERSPEECH*, pages 1517–1520. ISCA.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed Abdel-Wahab, Najmeh Sadoughi, and Emily Mower Provost. 2017. MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. *IEEE Trans. Affect. Comput.*, 8(1):67–80.

Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intell. Syst.*, 31(2):102–107.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis.

In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.*, 5(4):377–390.

Shayok Chakraborty, Vineeth Nallure Balasubramanian, Qian Sun, Sethuraman Panchanathan, and Jieping Ye. 2015. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(10):1945–1958.

Yi Chang, Zhao Ren, Thanh Tam Nguyen, Kun Qian, and Björn W. Schuller. 2023. Knowledge transfer for on-device speech emotion recognition with neural structured learning. In *Proceedings of ICASSP*, pages 1–5.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of SemEval@NAACL-HLT*, pages 39–48.

Aditi Chaudhary, Zaid Sheikh, Antonis Anastasopoulos, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Trans. Assoc. Comput. Linguistics*, 9:1–16.

Maximillian Chen and Zhou Yu. 2023. Pre-finetuning for few-shot emotional speech recognition. In *Proceedings of INTERSPEECH*, pages 3602–3606.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrusaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of ICMI*, pages 163–171.

Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. 2023a. Vesper: A compact and effective pretrained model for speech emotion recognition. *CoRR, abs/2307.10757*.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2023b. DST: deformable speech transformer for emotion recognition. In *Proceedings of ICASSP*, pages 1–5.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of ACL/IJCNLP*, pages 5904–5914.

Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. NNIME: the NTHU-NTUA chinese interactive multimodal emotion corpus. In *Proceedings of ACII*, pages 292–298.

Chloé Clavel, Ioana Vasilescu, Laurence Devillers, Gaël Richard, and Thibaut Ehrette. 2008. Feartype emotion recognition for audio-based vasilescu systems. *Speech Communication*, 50(6):487–503.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of ACL*, pages 32–42.

Silvia Corchs, Elisabetta Fersini, and Francesca Gasparini. 2019. Ensemble learning on visual and textual data for social image emotion classification. *Int. J. Mach. Learn. Cybern.*, 10(8):2057–2070.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. EMOVO corpus: an italian emotional speech database. In *Proceedings of LREC*, pages 3501–3504.

Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George N. Votsis, Stefanos D. Kollias, Winfried A. Fellenz, and John G. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.*, 18(1):32–80.

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of NAACL-HLT*, pages 5305–5316.

An Dang, Toan H. Vu, Le Dinh Nguyen, and Jia-Ching Wang. 2023. EMIX: A data augmentation method for speech emotion recognition. In *Proceedings of ICASSP*, pages 1–5.

Jianwu Dang, Aijun Li, Donna Erickson, Atsuo Suemitsu, Masato Akagi, Kyoko Sakuraba, Nobuaki Minematsu, and Keikichi Hirose. 2010. Comparison of emotion perception among different cultures. *Acoustical science and technology*, 31(6):394–402.

Mohammad Darwich, Shahrul Azman Mohd. Noah, Nazlia Omar, and Nurul Aida Osman. 2019. Corpus-based techniques for sentiment lexicon generation: A review. *J. Digit. Inf. Manag.*, 17(5):296.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Workshop on Multimodal Language (Challenge-HML)*, pages 1–7.

Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn W. Schuller. 2018. Semisupervised autoencoders for speech emotion recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(1):31–43.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of ICMI*, pages 423–426.

Vipula Dissanayake, Sachith Seneviratne, Hussel Suriyaarachchi, Elliott Wen, and Suranga Nanayakkara. 2022. Self-supervised representation fusion for speech and wearable based emotion recognition. In *Proceedings of INTERSPEECH*, pages 3598–3602.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL*, pages 1383–1392.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of EMNLP*, pages 7949–7962.

Hicham El Boukkouri. 2021. *Domain adaptation of word embeddings through the exploitation of in-domain corpora and knowledge bases*. Theses, Université Paris-Saclay.

Nelly Elsayed, Zag ElSayed, et al. 2022. Speech emotion recognition using supervised deep recurrent system for mental health monitoring. *CoRR,abs/2208.12812*.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Trans. Affect. Comput.*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of MM*, page 1459–1462.

Weiquan Fan, Xiangmin Xu, Xiaofen Xing, Weidong Chen, and Dongyan Huang. 2021. Lssed: A large-scale dataset and benchmark for speech emotion recognition. In *Proceedings of ICASSP*, pages 641–645.

Misbah Farooq, Fawad Hussain, Naveed Khan Baloch, Fawad Riasat Raja, Heejung Yu, and Yousaf Bin Zikria. 2020. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors*, 20(21):6008–6015.

Kavan Fatehi and Ayse Kucukyilmaz. 2023. LABERT: A Combination of Local Aggregation and Self-Supervised Speech Representation Learning for Detecting Informative Hidden Units in Low-Resource ASR Systems. In *Proceedings INTERSPEECH 2023*, pages 211–215.

Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *Proceedings of IJCNN*, pages 566–570.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of EMNLP*, pages 1615–1625.

GdR Oliveira Ferreira. 2022. Domain specific wav2vec 2.0 fine-tuning for the se&r 2022 challenge. *CoRR, abs/2207.14418*.

Elisabetta Fersini, Enza Messina, and Federico Alberto Pozzi. 2014. Sentiment analysis: Bayesian ensemble learning. *Decis. Support Syst.*, 68:26–38.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of NAACL*, pages 2591–2597.

Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, and Jianwu Dang. 2021. CONSK-GCN: conversational semantic- and knowledge-oriented graph convolutional network for multimodal emotion recognition. In *Proceedings of ICME*, pages 1–6.

Sadaoki Furui. 1986. Speaker-independent isolated word recognition based on emphasized spectral dynamics. In *Proceedings of ICASSP*, pages 1991–1994.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, pages 6894–6910.

Soumaya Gharsellaoui, Sid-Ahmed Selouani, and Mohammed Sidi Yakoub. 2019. Linear discriminant differential evolution for feature selection in emotional speech recognition. In *Proceedings of INTERSPEECH*, pages 3297–3301.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of EMNLP*, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of EMNLP-IJCNLP*, pages 154–164.

Ayoub Ghriss, Bo Yang, Viktor Rozgic, Elizabeth Shriberg, and Chao Wang. 2022. Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition. In *Proceedings of ICASSP*, pages 7347–7351.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron C. Courville, Mehdi Mirza, Benjamin Hamner, William Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Tudor Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63.

Eric Guizzo, Tillman Weyde, Simone Scardapane, and Danilo Comminiello. 2022. Learning speech emotion representations in the quaternion domain. *CoRR, abs/2204.02385*.

Xiaobao Guo, Adams Kong, Huan Zhou, Xianfeng Wang, and Min Wang. 2021. Unimodal and crossmodal refinement network for multimodal sequence fusion. In *Proceedings of EMNLP*, pages 9143–9153.

Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, and Dinesh Manocha. 2022. 3massiv: Multilingual, multimodal and multi-aspect dataset of social media short videos. In *Proceedings of CVPR*, pages 21032–21043.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, pages 8342–8360.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of NeurIPS*, pages 1024–1034.

Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Proceedings of INTERSPEECH*, pages 223–227.

Wei Han, Hui Chen, Alexander F. Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria. 2021a. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of ICMI*, pages 6–15.

Wei Han, Hui Chen, and Soujanya Poria. 2021b. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of EMNLP*, pages 9180–9192.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of NAACL*, pages 2122–2132.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of MM*, page 1122–1131.

Fatemeh Hemmatian and Mohammad Karim Sohrabi. 2019. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial intelligence review*, 52(3):1495–1545.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of LREC*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022a. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In *Proceedings of ICASSP*, pages 7037–7041.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of ACL*, pages 7042–7052.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022b. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of EMNLP*, pages 7837–7851.

Chenchen Huang, Wei Gong, Wenlong Fu, and Dongyu Feng. 2014a. A research of speech emotion recognition based on deep belief network and svm. *Mathematical Problems in Engineering*, 2014(1):749604.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017a. Densely connected convolutional networks. In *Proceedings of CVPR*, pages 2261–2269.

Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. 2018. Multimodal continuous emotion recognition with data augmentation using recurrent neural networks. In *Proceedings of AVEC@MM*, pages 57–64.

Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Zhengqi Wen, Minghao Yang, and Jiangyan Yi. 2017b. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In *Proceedings of ASRU*, pages 11–18.

Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014b. Speech emotion recognition using CNN. In *Proceedings of MM*, pages 801–804.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of EMNLP*, pages 7360–7370.

Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

Prayas Jain, Pranav Goel, Devang Kulshreshtha, and Kaushal Kumar Shukla. 2017. Prayas at emoint 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. In *Proceedings of WASSA@EMNLP*, pages 58–65.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of AAAI*, pages 8002–8009.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of NAACL*, pages 4148–4164.

Takeo Kanade, Ying-li Tian, and Jeffrey F. Cohn. 2000. Comprehensive database for facial expression analysis. In *Proceedings of FG*, pages 46–53.

Leo Katz. 1953. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of EMNLP*, pages 1160–1166.

Yeachan Kim and Bonggun Shin. 2022. In defense of core-set: A density-aware core-set selection for active learning. In *Proceedings of SIGKDD*, pages 804–812.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, pages 1–15.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR*, pages 1–14.

Jan Kocon. 2023. Deep emotions across languages: A novel approach for sentiment propagation in multilingual wordnets. In *Proceedings of ICDM*, pages 744–749.

Leonard Konle and Fotis Jannidis. 2020. Domain and task adaptive pretraining for language models. In *Proceedings of CHR*, volume 2723, pages 248–256.

Abhishek Kumar, Daisuke Kawahara, and Sadao Kurohashi. 2018. Knowledge-enriched two-layered attention network for sentiment analysis. In *Proceedings of NAACL-HLT*, pages 253–258.

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Björn W. Schuller. 2022. Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Trans. Affect. Comput.*, 13(2):992–1004.

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn W. Schuller. 2023. Survey of deep representation learning for speech emotion recognition. *IEEE Trans. Affect. Comput.*, 14(2):1634–1654.

Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps. 2018. Transfer learning for improving speech emotion classification accuracy. In *Proceedings of INTERSPEECH*, pages 257–261.

Chandrashekhar Lavania, Sanjiv Das, Xin Huang, and Kyu Jeong Han. 2023. Utility-preserving privacy-enabled speech embeddings for emotion detection. In *Proceedings of INTERSPEECH*, pages 3612–3616.

L.Chen and A.Rudnicky. 2022. Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition. In *Proceedings of ICASSP*, pages 1–5.

Hung Le, Nancy Chen, and Steven Hoi. 2022. Multimodal dialogue state tracking. In *Proceedings of NAACL*, pages 3394–3415.

Yann LeCun and Yoshua Bengio. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Yann LeCun, Yoshua Bengio, et al. 2015. Deep learning. *Nat.*, 521(7553):436–444.

Howard Levene et al. 1960. Contributions to probability and statistics. *Essays in honor of Harold Hotelling*, 278:292.

Dongyuan Li, Qiang Lin, and Xiaoke Ma. 2021a. Identification of dynamic community in temporal network via joint learning graph representation and nonnegative matrix factorization. *Neurocomputing*, 435:77–90.

Dongyuan Li and Xiaoke Ma. 2019. Nonnegative matrix factorization for dynamic modules in cancer attribute temporal networks. In *Proceedings of BIBM*, pages 202–206.

Dongyuan Li, Xiaoke Ma, and Maoguo Gong. 2021b. Joint learning of feature extraction and clustering for large-scale temporal networks. *IEEE Transactions on Cybernetics*, 53(3):1653–1666.

Dongyuan Li, Shiyin Tan, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023a. Temporal and topological augmentation-based cross-view contrastive learning model for temporal link prediction. In *Proceedings of CIKM*, pages 4059–4063.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023b. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In *Proceedings of EMNLP*, pages 16051–16069.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. 2024. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Dongyuan Li, Jingyi You, Kotaro Funakoshi, and Manabu Okumura. 2022a. A-TIP: attribute-aware text infilling via pre-trained language model. In *Proceedings of COLING*, pages 5857–5869. International Committee on Computational Linguistics.

Dongyuan Li, Shuyao Zhang, and Xiaoke Ma. 2022b. Dynamic module detection in temporal attributed networks of cancers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(4):2219–2230.

Sha Li, Madhi Namazifar, Di Jin, MOHIT BANSAL, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022c. Enhanced knowledge selection for grounded dialogues via document semantic graphs. In *NAACL 2022*.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022d. Contrast and generation make BART a good dialogue emotion recognizer. In *Proceedings of AAAI*, pages 11002–11010.

Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong. 2022e. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of AAAI*, pages 4541–4549.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022f. CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of NAACL*, pages 2282–2294.

Zheng Lian, Bin Liu, and Jianhua Tao. 2023. SMIN: semi-supervised multi-modal interaction network for conversational emotion recognition. *IEEE Trans. Affect. Comput.*, 14(3):2415–2429.

Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. 2020. Conversational emotion recognition using self-attention mechanisms and graph neural networks. In *Proceedings of INTERSPEECH*, pages 2347–2351.

Sheng Liang, Mengjie Zhao, and Hinrich Schuetze. 2022. Modular and parameter-efficient multimodal fusion with prompting. In *Findings of ACL*, pages 2976–2985.

Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. 2021. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *Proceedings of ICCV*, pages 8128–8136.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2023. Summary-oriented vision modeling for multimodal abstractive summarization. In *Proceedings of ACL*, pages 2934–2951.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of ACL*, pages 2149–2159.

Jun Liu, Hongxia Wang, and Yanjun Feng. 2021. An end-to-end deep model with discriminative facial features for facial expression recognition. *IEEE Access*, 9:12158–12166.

Mingyi Liu, Zhiying Tu, Tong Zhang, Tonghua Su, Xiaofei Xu, and Zhongjie Wang. 2022. LTP: A new active learning strategy for crf-based named entity recognition. *Neural Process. Lett.*, 54(3):2433–2454.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multi-stage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of EMNLP*, pages 1834–1845.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR, abs/1907.11692*.

Reza Lotfian and Carlos Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affect. Comput.*, 10(4):471–483.

Han Lu, Xiahai Zhuang, and Qiang Luo. 2024. A brain-inspired way of reducing the network complexity via concept-regularized coding for emotion recognition. In *Proceedings of AAAI*, pages 556–564.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of CVPR*, pages 2554–2562.

Michael J. Lyons, Miyuki Kamachi, and Jiro Gyoba. 2020. Coding facial expressions with gabor wavelets. *CoRR*, abs/2009.05938.

Yaxiong Ma, Yixue Hao, Min Chen, Jincai Chen, Ping Lu, and Andrej Košir. 2019. Audio-visual emotion fusion (avef): A deep efficient weighted approach. *Information Fusion*, 46:184–192.

Zohreh Madhoushi, Abdul Razak Hamdan, and Suhaila Zainudin. 2015. Sentiment analysis techniques in recent works. In *Proceddings of SAI*, pages 288–291.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *Proceedings of AAAI*, pages 6818–6825.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-SENA: An integrated platform for multimodal sentiment analysis. In *Proceedings of ACL*, pages 204–213.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of EMNLP*, pages 650–663.

Gary McKeown, Michel François Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.*, 3(1):5–17.

M.Dredze and K.Crammer. 2008. Active learning with confidence. In *Proceedings of ACL*, pages 233–236.

Hardik Meisheri and Lipika Dey. 2018. TCS research at semeval-2018 task 1: Learning robust representations using multi-attention architecture. In *Proceedings of SemEval@NAACL-HLT*, pages 291–299.

Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. 2010. Decision level combination of multiple modalities for recognition of emotional expression. In *Proceedings of ICASSP*, pages 2462–2465.

Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proceedings of ICASSP*, pages 2227–2231.

Omar Mohamed and Salah A. Aly. 2021. Arabic speech emotion recognition employing wav2vec2.0 and hubert based on BAVED dataset. *CoRR, abs/2110.04425*.

Sina Mohseni, Niloofar Zarei, and Saba Ramazani. 2014. Facial expression recognition using anatomy based facial graph. In *Proceedings of SMC*, pages 3715–3719.

Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech emotion recognition using self-supervised features. In *Proceedings of ICASSP*, pages 6922–6926.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of ICMI*, pages 169–176.

M.Yuan, H.Lin, and JB.Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of EMNLP*, pages 7935–7948.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. In *Proceedings of NeurIPS*, pages 14200–14213.

Shahla Nemati, Reza Rohani, Mohammad Ehsan Basiri, Moloud Abdar, Neil Y. Yen, and Vladimir Makarenkov. 2019. A hybrid latent space data fusion method for multimodal emotion recognition. *IEEE Access*, 7:172948–172964.

Hai Duong Nguyen, Sun-Hee Kim, Guee-Sang Lee, Hyung-Jeong Yang, In Seop Na, and Soo-Hyung Kim. 2022. Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks. *IEEE Trans. Affect. Comput.*, 13(1):226–237.

Tuan-Linh Nguyen, Swathi Kavuri, and Minho Lee. 2019. A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips. *Neural Networks*, 118:208–219.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. e-health CSIRO at radsum23: Adapting a chest x-ray report generator to multimodal radiology report summarisation. In *The 22nd Workshop on BioNLP@ACL*, pages 545–549.

Weizhi Nie, Minjie Ren, Jie Nie, and Sicheng Zhao. 2021. C-GCN: correlation based graph convolutional network for audio-video emotion recognition. *IEEE Trans. Multim.*, 23:3793–3804.

Kosuke Nishida, Kyosuke Nishida, and Sen Yoshida. 2021. Task-adaptive pre-training of language models with word embedding regularization. In *Proceedings of ACL*, pages 4546–4553.

M.Karami OM.Nezami, PJ.Lou. 2019. Shemo: a large-scale validated database for persian speech emotion detection. *Language Resources and Evaluation*, 53(1):1–16.

Juan D. S. Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli, Patrick Cardinal, and Alessandro L. Koerich. 2019. Multimodal fusion with deep neural networks for audio-video emotion recognition. *CoRR*, abs/1907.03196.

Timothy Ossowski and Junjie Hu. 2023. Retrieving multimodal prompts for generative visual question answering. In *Findings of the ACL*.

Sarala Padi, Seyed Omid Sadjadi, et al. 2021. Improved speech emotion recognition using transfer learning and spectrogram augmentation. In *Proceedings of ICMI*, pages 645–652.

Georgios Paraskevopoulos, Efthymios Tzinis, Nikolaos Ellinas, Theodoros Giannakopoulos, and Alexandros Potamianos. 2019. Unsupervised low-rank representations for speech emotion recognition. In *Proceedings of INTERSPEECH*, pages 939–943.

Sarah Partan and Peter Marler. 1999. Communication goes multimodal. *Science*, 283(5406):1272–1273.

Srinivas Parthasarathy and Carlos Busso. 2018. Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes. In *Proceedings of INTERSPEECH*, pages 3698–3702.

Yifan Peng, Yui Sudo, Shakeel Muhammad, and Shinji Watanabe. 2023. DPHuBERT: Joint Distillation and Pruning of Self-Supervised Speech Models. In *Proceedings of INTERSPEECH*, pages 62–66.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Proceedings of INTERSPEECH*, pages 3400–3404. ISCA.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of ACL*, pages 973–982.

Isidoros Perikos and Ioannis Hatzilygeroudis. 2016. Recognizing emotions in text using ensemble of classifiers. *Eng. Appl. Artif. Intell.*, 51:191–201.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of ICCV*, pages 9616–9625.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of ACL*, pages 873–883.

Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59.

Soujanya Poria, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2015. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, 63:104–116.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL*, pages 527–536.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander F. Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations. *Cogn. Comput.*, 13(5):1317–1332.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard H. Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Tuomas Puoliväli, Satu Palva, and J. Matias Palva. 2020. Influence of multiple hypothesis testing on reproducibility in neuroimaging research: A simulation study and python-based software. *Journal of Neuroscience Methods*, 337:108654.

Preeth Raguraman, Mohan Ramasundaram, and Midhula Vijayan. 2019. Librosa based assessment tool for music information retrieval systems. In *Proceedings of MIPR*, pages 109–114.

Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Mohammed E. Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of ACL*, pages 2359–2369.

Vandana Rajan, Alessio Brutti, and Andrea Cavallaro. 2022. Is cross-attention preferable to self-attention for multi-modal emotion recognition? In *Proceedings of ICASSP*, pages 4693–4697.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.

Zhao Ren, Thanh Tam Nguyen, Yi Chang, and Björn W. Schuller. 2022. Fast yet effective speech emotion recognition with self-distillation. *CoRR, abs/2210.14636*.

Fabien Ringeval, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proceedings of FG*, pages 1–8.

R.Nicholas and M.Andrew. 2001. Toward optimal active learning through monte carlo estimation of error reduction. *Proceedings of ICML*, 2:441–448.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proceedings of AAAI*, pages 1611–1618.

Guy Rotman and Roi Reichart. 2022. Multi-task active learning for pre-trained transformer-based models. *Trans. Assoc. Comput. Linguistics*, 10:1209–1228.

Vin Sachidananda, Jason S. Kessler, and Yi'an Lai. 2021. Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of EMNLP*, pages 155–165.

Saurabh Sahu, Rahul Gupta, and Carol Y. Espy-Wilson. 2022. Modeling feature representations for affective speech using generative adversarial networks. *IEEE Trans. Affect. Comput.*, 13(2):1098–1110.

Saurabh Sahu, Rahul Gupta, Ganesh Sivaraman, Wael AbdAlmageed, and Carol Y. Espy-Wilson. 2017. Adversarial auto-encoders for speech based emotion recognition. In *Proceedings of ISCA*, pages 1243–1247.

Rachid Sammouda and Ali El-Zaart. 2021. An optimized approach for prostate image segmentation using k-means clustering algorithm with elbow method. *Comput. Intell. Neurosci.*, 2021:4553832:1–4553832:13.

Jennifer Santoso, Takeshi Yamada, Kenkichi Ishizuka, Taiichi Hashimoto, and Shoji Makino. 2022. Speech emotion recognition based on self-attention weight correction for acoustic and text features. *IEEE Access*, 10:115732–115743.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proceedings of INTERSPEECH*, pages 3465–3469.

Björn W. Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.*, 53(9-10):1062–1087.

Brian B Schultz. 1985. Levene's test for relative variation. *Systematic Zoology*, 34(4):449–456.

Shiv Shankar. 2022. Multimodal fusion via cortical network inspired losses. In *Proceedings of ACL*, pages 1167–1178.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality. *Biometrika*, 52(3/4):591–611.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of ICML*, volume 80, pages 4603–4611.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of AAAI*, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of ACL/IJCNLP*, pages 1551–1560.

Dongming Sheng, Dong Wang, Ying Shen, Haitao Zheng, and Haozhuang Liu. 2020. Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition. In *Proceedings of COLING*, pages 4153–4163.

Aman Shenoy, Ashish Sardana, and et al. 2020. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Workshop on Multimodal Language (Challenge-HML)*, pages 19–28.

Bei Shi, Zihao Fu, Lidong Bing, and Wai Lam. 2018. Learning domain-sensitive and sentiment-aware word embeddings. In *Proceedings of ACL*, pages 2494–2504.

Tao Shi and Shao-Lun Huang. 2023. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of ACL*, pages 14752–14766.

Kimiko Shimoda, Michael Argyle, and Pio Ricci Bitti. 1978. The intercultural recognition of emotional expressions by three national racial groups: English, italian and japanese. *European Journal of Social Psychology*, 8(2):169–179.

Apoorva Singh, Soumyodeep Dey, Anamitra Singha, and Sriparna Saha. 2022. Sentiment and emotion-aware multi-modal complaint identification. In *Proceedings of AAAI*, pages 12163–12171.

Sundararajan Srinivasan, Zhaocheng Huang, and Katrin Kirchhoff. 2022. Representation learning through cross-modal conditional teacher-student training for speech emotion recognition. In *Proceedings of ICASSP*, pages 6442–6446.

SR.Livingstone and FA.Russo. 2018. The ryerson audio-visual database of emotional speech and song. *PLOS ONE*, 13(5):1–35.

Haiyang Su, Bin Liu, Jianhua Tao, Yongfeng Dong, Jian Huang, Zheng Lian, and Leichao Song. 2020. An improved multimodal dimension emotion recognition based on different fusion methods. In *Proceedings of ICSP*, volume 1, pages 257–261.

Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. 2020. Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 27–34.

Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor detection on social media with graph adversarial contrastive learning. In *Proceedings of WWW*, pages 2789–2797.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Zeinab Sadat Taghavi, Ali Satvaty, and Hossein Sameti. 2023. A change of heart: Improving speech emotion recognition through speech-to-text modality conversion. In *Proceedings of ICLR*.

Shiyin Tan, Jingyi You, and Dongyuan Li. 2022. Temporality- and frequency-aware graph contrastive learning for temporal network. In *Proceedings of CIKM*, pages 1878–1888.

Jianhua Tao, Jian Huang, Ya Li, Zheng Lian, and Mingyue Niu. 2019. Semi-supervised ladder networks for speech emotion recognition. *Int. J. Autom. Comput.*, 16(4):437–448.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019a. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of ACL*, pages 6558–6569.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019b. Learning factorized multimodal representations. In *Proceedings of ICLR*.

Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of EMNLP*, pages 1823–1833.

Turker Tuncer, Sengul Dogan, and U Rajendra Acharya. 2021. Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. *Knowledge-Based Systems*, 211:106547.

Panagiotis Tzirakis, Jiaxin Chen, Stefanos Zafeiriou, and Björn W. Schuller. 2021. End-to-end multimodal affect recognition in real-world environments. *Inf. Fusion*, 68:46–53.

Md Azher Uddin, Joolekha Bibi Joolee, and Kyung-Ah Sohn. 2021. Dynamic facial expression understanding using deep spatiotemporal LDSP on spark. *IEEE Access*, 9:16866–16877.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of AAAI*, pages 173–177.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2023. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13.

Elaine Walker, Samuel Marwit, and Eugene Emory. 1980. A cross-sectional study of emotion recognition in schizophrenics. *Journal of Abnormal Psychology*, 89(3):428.

Hu Wang. 2018. Renn: Rule-embedded neural networks. In *Proceedings of ICPR*, pages 824–829.

Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. 2020a. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.*, 29:4057–4069.

Qianning Wang, Chenglin Wang, Zhixin Lai, and Yucheng Zhou. 2024a. Insectmamba: Insect pest classification with state space model. *CoRR, abs/2404.03611*.

Tana Wang, Yaqing Hou, Dongsheng Zhou, and Qiang Zhang. 2021. A contextual attention network for multimodal emotion recognition in conversation. In *Proc. of IJCNN*, pages 1–7.

Xiaohua Wang, Jianqiao Gong, Min Hu, Yu Gu, and Fuji Ren. 2020b. LAUN improved stargan for facial emotion recognition. *IEEE Access*, 8:161509–161518.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020c. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of SIGDIAL*, pages 186–195.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of AAAI*, pages 7216–7223.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. 2022a. Multimodal token fusion for vision transformers. In *Proceedings of CVPR*, pages 12176–12185.

Yiming Wang, Xinghui Dong, Gongfa Li, Junyu Dong, and Hui Yu. 2022b. Cascade regression-based face frontalization for dynamic facial expression analysis. *Cogn. Comput.*, 14(5):1571–1584.

Yusong Wang, Dongyuan Li, Kotaro Funakoshi, and Manabu Okumura. 2023. Emp: emotion-guided multi-modal fusion and contrastive learning for personality traits recognition. In *Proceedings of ICMR*, pages 243–252.

Yusong Wang, Dongyuan Li, and Jialun Shen. 2024b. Inter-modality and intra-sample alignment for multi-modal emotion recognition. In *Proceedings of ICASSP*, pages 8301–8305.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022c. N24news: A new dataset for multimodal news classification. In *Proceedings of LREC*, pages 6768–6775.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of MM*, pages 1437–1445.

Xin-Cheng Wen, Jia-Xin Ye, Yan Luo, Yong Xu, Xuan-Ze Wang, Chang-Li Wu, and Kun-Hong Liu. 2022. Ctl-mtnet: A novel capsnet and transfer learning-based mixed task net

for single-corpus and cross-corpus speech emotion recognition. In *Proceedings of IJCAI*, pages 2305–2311.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of Challenge-HML*, pages 11–19.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn W. Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intell. Syst.*, 28(3):46–53.

Ting Wu, Junjie Peng, Wenqiang Zhang, Huiran Zhang, Shuhua Tan, Fen Yi, Chuanshuai Ma, and Yansong Huang. 2022. Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowl. Based Syst.*, 235:107676.

Yang Wu, Pengwei Zhan, Yunjian Zhang, LiMing Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the ACL/IJCNLP*, pages 2560–2569.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of ICASSP*, pages 1–5.

Yangyang Xia, Li-Wei Chen, Alexander Rudnicky, and Richard M Stern. 2021. Temporal context in speech emotion recognition. In *Proceedings of INTERSPEECH*, pages 3370–3374.

Detai Xin, Shinnosuke Takamichi, Ai Morimatsu, and Hiroshi Saruwatari. 2023. Laughter synthesis using pseudo phonetic tokens with a large-scale in-the-wild laughter corpus. *CoRR, abs/2305.12442*.

Dongkuan Xu, Wei Cheng, Dongsheng Luo, Haifeng Chen, and Xiang Zhang. 2021. Infogcl: Information-aware graph contrastive learning. In *Proceedings of NeurIPS*, pages 30414–30425.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2vec: Learning generalized emotion representation by multi-task training. In *Proceedings of WASSA@EMNLP*, pages 292–298.

Yan Xu, Fuming Sun, and Xue Zhang. 2013. Literature survey of active learning in multimedia annotation and retrieval. In *Proceedings of ICIMCS*, pages 237–242.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. MTAG: modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of NAACL-HLT*, pages 1009–1021.

Yiming Yang, Xin Liu, and et al. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR*, page 42–49.

Jia-Xin Ye, Xin-Cheng Wen, Xuan-Ze Wang, Yong Xu, Yan Luo, Chang-Li Wu, Li-Yan Chen, and Kun-Hong Liu. 2022. Gm-tcnet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition. *Speech Communication*, 145:21–35.

Jiaxin Ye, Xin-Cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. 2023. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *Proceedings of ICASSP*, pages 1–5.

Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. 2006. A 3d facial expression database for facial behavior research. In *Proceedings of FGR*, pages 211–216.

Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP*, pages 316–321.

Jingyi You, Dongyuan Li, Manabu Okumura, and Kenji Suzuki. 2022. JPG - jointly learn to align: Automated disease prediction and radiology report generation. In *Proceedings of COLING*, pages 5989–6001.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. In *Proceedings of NeurIPS*.

Mingjing Yu, Huicheng Zheng, Zhifeng Peng, Jiayu Dong, and Heran Du. 2020. Facial expression recognition based on a multi-task global-local network. *Pattern Recognit. Lett.*, 131:166–171.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of AAAI*, pages 10790–10797.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of EMNLP*, pages 7935–7948.

Robert H Zabel. 1979. Recognition of emotions in facial expressions by emotionally disturbed and nondisturbed children. *Psychology in the Schools*, 16(1):119–126.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of EMNLP*, pages 1103–1114.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of AAAI*, pages 5634–5641.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of ACL*, pages 2236–2246.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018c. Multi-attention recurrent network for human communication comprehension. In *Proceedings of AAAI*, pages 5642–5649.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016a. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259.

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

Jiaqi Zeng and Pengtao Xie. 2021. Contrastive self-supervised learning for graph classification. In *Proceedings of AAAI*, pages 10824–10832.

Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. 2017. Baum-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313.

Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021a. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *Proceedings of AAAI*, pages 14338–14346.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of IJCAI*, pages 5415–5421.

Leihan Zhang and Le Zhang. 2020. An ensemble deep active learning method for intent classification. In *Proceedings of CSAI*, page 107–111.

Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. 2018a. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multim.*, 20(6):1576–1590.

Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2018b. Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 28(10):3030–3043.

Shiqing Zhang, Xiaoming Zhao, and Qi Tian. 2022a. Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. *IEEE Trans. Affect. Comput.*, 13(2):680–688.

Xi Zhang, Feifei Zhang, and Changsheng Xu. 2022b. Joint expression synthesis and representation learning for facial expression recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 32(3):1681–1695.

Xiaoliang Zhang, Yulin He, Yi Jin, Honglian Qin, Muhammad Azhar, and Joshua Zhexue Huang. 2020. A robust k-means clustering algorithm based on observation point mechanism. *Complex.*, 2020:1–11.

Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, and Irwin King. 2022c. COSTA: covariance-preserving feature augmentation for graph contrastive learning. In *Proceedings of KDD*, pages 1–18.

Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura. 2021b. A language model-based generative classifier for sentence-level discourse parsing. In *Proceedings of EMNLP*, pages 2432–2446.

Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura. 2023. Bidirectional transformer reranker for grammatical error correction. In *Findings of ACL*, pages 3801–3825.

Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura. 2024. Bidirectional transformer reranker for grammatical error correction. *Journal of Natural Language Processing*, 31(1):3–46.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022d. A survey of active learning for natural language processing. In *Proceedings of EMNLP*, pages 6166–6190.

Huan Zhao, Yufeng Xiao, and Zixing Zhang. 2020. Robust semisupervised generative adversarial networks for speech emotion recognition via distribution smoothness. *IEEE Access*, 8:106889–106900.

Rui Zhao, Tianshan Liu, Zixun Huang, Daniel Pak-Kong Lun, and Kin-Man Lam. 2023. Geometry-aware facial expression recognition via attentive graph convolutional networks. *IEEE Trans. Affect. Comput.*, 14(2):1159–1174.

Wenjing Zhu and Xiang Li. 2022. Speech emotion recognition with global-aware fusion on multi-scale feature representation. In *Proceedings of ICASSP*, pages 6437–6441.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. Deep Graph Contrastive Representation Learning. In *Proceedings of ICML Workshop on GRLB*.

Cairong Zou, Xinran Zhang, Cheng Zha, and Li Zhao. 2016. A novel DBN feature fusion model for cross-corpus speech emotion recognition. *J. Electr. Comput. Eng.*, 2016:7437860:1–7437860:11.

Selected Publications

Publications Related to Doctoral Thesis

Journal Papers

[1] Active Learning with Task Adaptation Pre-training for Speech Emotion Recognition. **Dongyuan Li**, Ying Zhang, Yusong Wang, Funakoshi Kataro, Manabu Okumura. Journal of Natural Language Processing (**JNLP**), 2024.

Conference Papers

[1] After: Active learning based fine-tuning framework for speech emotion recognition. **Dongyuan Li**, Yusong Wang, Kotaro Funakoshi, Manabu Okumura. IEEE Automatic Speech Recognition and Understanding Workshop (**IEEE ASRU**), 2023.

[2] Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition. **Dongyuan Li**, Yusong Wang, Kotaro Funakoshi, Manabu Okumura. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (**EMNLP**), 2023.

Other Selected Publications (* means equal contribution)

Journal Papers

[1] A Survey on Deep Active Learning: Recent Advances and New Frontiers. **Dongyuan Li***, Zhen Wang*, Yankai Chen, Renhe Jiang, Weiping Ding, Manabu Okumura. IEEE

Transactions on Neural Networks and Learning Systems, (**IEEE TNNLS**), 2024.

[2] Plug-and-Play Attribute-Aware Text Infilling via A New Attention Mechanism and Two-Level Positional Encoding. **Dongyuan Li**, Kotaro Funakoshi, Manabu Okumura. Journal of Natural Language Processing, (**JNLP**), 2023.

[3] Joint learning-based heterogeneous graph attention network for timeline summarization. Jingyi You, **Dongyuan Li**, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura. Journal of Natural Language Processing (**JNLP**), 2023.

[4] Joint learning of feature extraction and clustering for large-scale temporal networks. **Dongyuan Li**, Xiaoke Ma, Maoguo Gong. IEEE Transactions on Cybernetics, (**IEEE TCBY**), 2023.

[5] Detecting dynamic community by fusing network embedding and nonnegative matrix factorization. **Dongyuan Li**, Xiaoxiong Zhong, Zengfa Dou, Maoguo Gong, Xiaoke Ma. Knowledge-Based Systems, (**KBS**), 2022.

[6] Identification of dynamic community in temporal network via joint learning graph representation and nonnegative matrix factorization. **Dongyuan Li**, Qiang Lin, Xiaoke Ma. **Neurocomputing**, 2022.

[7] Dynamic module detection in temporal attributed networks of cancers. **Dongyuan Li**, Shuyao Zhang, Xiaoke Ma. IEEE/ACM Transactions on Computational Biology and Bioinformatics, (**IEEE TCBB**), 2022.

Conference papers

[1] Community-Invariant Graph Contrastive Learning. Shiyin Tan*, **Dongyuan Li***, Renhe Jiang, Ying Zhang, Manabu Okumura. The Forty-first International Conference on Machine Learning, (**ICML**), 2024.

[2] Multimodal Graph-Based Audio-Visual Event Localization. Zhen Wang*, **Dongyuan Li***, Manabu Okumura. IEEE International Conference on Acoustics, Speech and Signal Processing (**IEEE ICASSP**), 2024.

[3] Temporal and Topological Augmentation-based Cross-view Contrastive Learning Model for Temporal Link Prediction. **Dongyuan Li**, Shiyin Tan, Yusong Wang, Kotaro Funakoshi, Manabu Okumura. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (**ACM CIKM**), 2023.

[4] A-TIP: attribute-aware text infilling via pre-trained language model. **Dongyuan Li**, Jingyi You, Kotaro Funakoshi, Manabu Okumura. Proceedings of the 29th International Conference on Computational Linguistics, (**COLING**), 2022.

[5] Emp: Emotion-guided multi-modal fusion and contrastive learning for personality traits recognition. Yusong Wang, **Dongyuan Li**, Kotaro Funakoshi, Manabu Okumura. Proceedings of the 2023 ACM International Conference on Multimedia Retrieval (**ICMR**), 2023.

[6] Jpg-jointly learn to align: Automated disease prediction and radiology report generation. Jingyi You, **Dongyuan Li**, Manabu Okumura, Kenji Suzuki. Proceedings of the 29th international conference on computational linguistics, (**COLING**), 2022.

[7] Joint learning-based Heterogeneous Graph Attention Network for Timeline Summarization. Jingyi You, **Dongyuan Li**, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (**NAACL**), 2022.