Journal of Artificial Intelligence Research 79 (2024) 825-863 Submitted 08/2023; published 03/2024

Multi-Modal Attentive Prompt Learning for Few-shot Emotion Recognition in Conversations

Xingwei Liang XINGWEI.LIANG@GMAIL.COM
Geng Tu TUGENG0313@GMAIL.COM
Jiachen Du JACOBVAN199165@GMAIL.COM
Harbin Institute of Technology, Shenzhen, P.R.China, 518055

Ruifeng Xu XURUIFENG@HIT.EDU.CN
Harbin Institute of Technology, Shenzhen, P.R.China, 518055
Peng Cheng Laboratory, Shenzhen, China
Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies

Abstract

Emotion recognition in conversations (ERC) has emerged as an important research area in Natural Language Processing and Affective Computing, focusing on accurately identifying emotions within the conversational utterance. Conventional approaches typically rely on labeled training samples for fine-tuning pre-trained language models (PLMs) to enhance classification performance. However, the limited availability of labeled data in real-world scenarios poses a significant challenge, potentially resulting in diminished model performance. In response to this challenge, we present the Multi-modal Attentive Prompt (MAP) learning framework, tailored specifically for few-shot emotion recognition in conversations. The MAP framework consists of four integral modules: multi-modal feature extraction for the sequential embedding of text, visual, and acoustic inputs; a multi-modal prompt generation module that creates six manually-designed multi-modal prompts; an attention mechanism for prompt aggregation; and an emotion inference module for emotion prediction. To evaluate our proposed model's efficacy, we conducted extensive experiments on two widely recognized benchmark datasets, MELD and IEMOCAP. Our results demonstrate that the MAP framework outperforms state-of-the-art ERC models, yielding notable improvements of 3.5% and 0.4% in micro F1 scores. These findings highlight the MAP learning framework's ability to effectively address the challenge of limited labeled data in emotion recognition, offering a promising strategy for improving ERC model performance.

1. Introduction

Nowadays, a growing number of users are engaging in multi-party conversations and expressing their opinions on social media platforms (Gluz & Jaques, 2017; Zhang et al., 2019). This has resulted in the generation of millions of conversation pages daily, which serve as valuable resources for understanding public sentiments and emotions. As a result, Emotion Recognition in Conversations (ERC) has emerged as an important research area in the natural language processing (NLP) community in recent years, owing to its numerous potential applications (Zhang et al., 2021b; Huddar et al., 2019). Here are some examples: (1) ERC can be applied in mental health domains to analyze and understand emotional states during therapeutic conversations or in online support groups (Fei et al., 2020); (2) it can be utilized in customer service interactions to gauge customer satisfaction, detect frustration and dissatisfaction, and enable personalized responses (Han et al., 2020); (3) it can be integrated into social robots or virtual assistants to enhance their ability to understand and respond to human emotions (Spezialetti et al., 2020); (4) it can be utilized in market research and advertising to analyze consumer sentiments and emotional responses to products, services, or advertisements (Le & Vea, 2016; Ribeiro et al., 2017). This information can guide marketing strategies, content creation, and product development.

Traditional emotion recognition (ER) focuses on identifying emotions from individual samples, such as text documents, images, or audio recordings. It aims to classify the emotional state or expression present in a single instance without considering the contextual information or interaction dynamics. The proposed ER approaches to multi-modal emotion recognition primarily focus on feature alignment and fusion. For example, Xiao et al. (2020) presented two domain adaptation methods, the generalized domain adversarial neural network (GDANN) and the class-aligned GDANN (CGDANN), to learn generalized domain-invariant representations for emotion recognition. Zhang et al. (2023) presented a multi-modal multi-task interactive graph attention network, termed M3GAT, to simultaneously solve the problems of multi-modal fusion and multi-task fusion. The experimental results proved the effectiveness.

In contrast, ERC specifically targets the identification of emotions within the conversational utterances. It takes into account the dynamic nature of conversations, where emotions can be influenced by the dialogue context, speaker interactions, and the overall discourse. This task aims to capture and interpret the emotional dynamics that emerge during conversations and understand how emotions evolve and are expressed throughout the dialogue. Unlike traditional emotion recognition, which often analyzes isolated instances, emotion recognition in conversations involves a broader analysis of the dialogue structure, speaker interactions, turn-taking, and the exchange of emotional information between participants. This contextual understanding is essential for accurately identifying emotions in conversational settings. The proposed ERC approaches place greater emphasis on contextual modeling. Since conversations involve multiple interacting participants and the temporal evolution of dialogue, methods for conversation emotion recognition consider the dialogue history and mutual influence among participants. This contextual modeling helps understand the dynamics of emotion changes, empathy, and the expression of emotions (Song et al., 2018a; Zhang et al., 2018). In spite of their success, recent studies have brought to light a significant limitation of Emotion Recognition in Conversations (ERC) models: their performance is greatly reliant on the availability of an ample supply of labeled samples. Nevertheless, owing to the intricacy and subjectivity involved in emotional understanding, the creation and annotation of a large-scale, high-quality ERC dataset pose a challenging undertaking with substantial annotation and time-related costs. Currently, only a few labeled multi-modal ERC datasets are in existence, such as MELD (Poria et al., 2019)

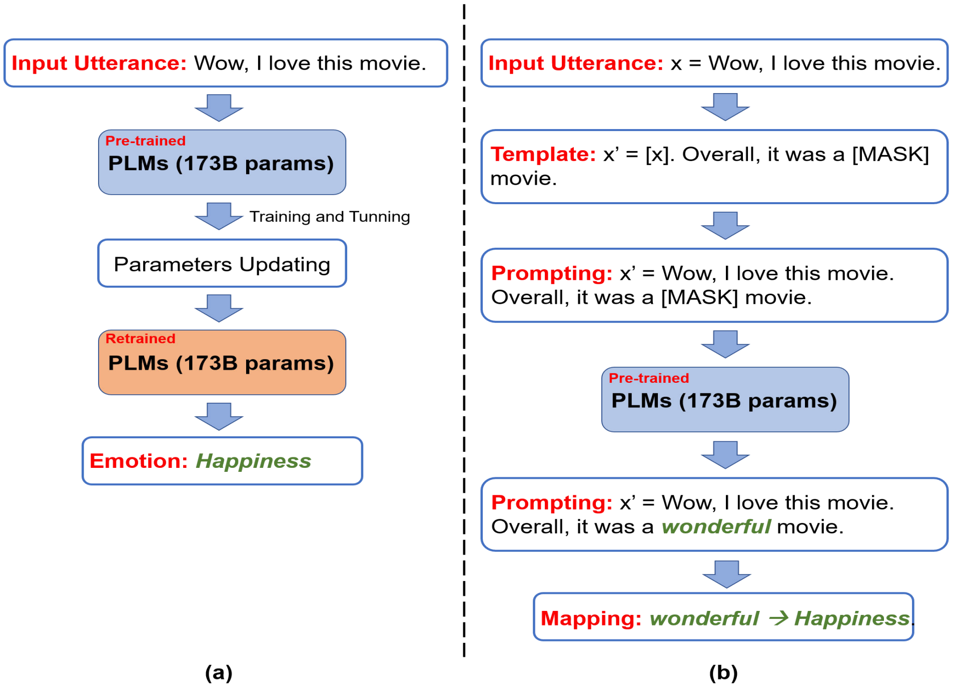
Input Utterance: Wow, I love this movie.
Pre-trained
PLMs (173B params)
Training and Tunning
Parameters Updating
Retrained
PLMs (173B params)
Emotion: Happiness
(a)
Input Utterance: x = Wow, I love this movie.
Template: x' = [x]. Overall, it was a [MASK] movie.
Prompting: x' = Wow, I love this movie. Overall, it was a [MASK] movie.
Pre-trained
PLMs (173B params)
Prompting: x' = Wow, I love this movie. Overall, it was a wonderful movie.
Mapping: wonderful → Happiness.
(b)

Figure 1: (a) The fine-tuning paradigm. (b) The prompt learning paradigm.

and IEMOCAP (Busso et al., 2008). IEMOCAP comprises approximately 12 hours of audiovisual conversations, while MELD encompasses 1,433 multi-party conversations, totaling around 33 hours of conversations. Despite the ongoing enlargement of Pre-trained Language Models (PLMs), the limited size of these datasets impedes the potential advancement of ERC tasks.

To address this issue, we propose the prompt learning paradigm for few-shot ERC. Few-shot ERC significantly differs from traditional ERC, where the few-shot ERC involves less training samples. Recently, with the introduction of GPT-3 (Brown et al., 2020), which has 175 billion parameters, a new PLM-based paradigm called ”prompt learning” has emerged. This approach enables accurate predictions by designing novel task descriptions (i.e., prompt), without updating any parameters of GPT-3. Unlike the “fine-tuning” paradigm, it does not require a large training dataset, making it more suitable for few-shot tasks, as illustrated in Fig. 1. One widely used framework in prompt learning involves designing a predefined template, such as *this movie is [MASK]* for text classification. In this case, the classification results depend on the probabilities of the predefined label words, such as “fantastic” or “terrible”, within the masked PLMs, namely the cloze problem. Note that the cloze problem refers to a type of language modeling task where a portion of the text, typically a word or a few words, is masked or removed, and the language model is required to predict the missing content based on the surrounding context. It is utilized as a way to train and fine-tune language models. By creating prompts with masked portions, the model is prompted to generate the missing content, which helps in shaping its language generation capabilities.

Prompt learning has quickly attracted attention of both academia and industry. An increasing number of prompt learning based approaches have been proposed to address various few-shot NLP tasks. For example, Wu & Shi (2022a) introduced an adversarial soft prompt tuning method (AdSPT) to improve cross-domain text sentiment analysis, achieving new state-of-the-art results. AdSPT is a technique that allows for fine-grained control over the behavior of language models by leveraging prompts. It involves tuning the prompts or instructions given to language models to improve their performance and control their outputs. The goal is to find an optimal set of prompts that guide the language model towards generating desired responses while minimizing the generation of unwanted or biased outputs. Xu et al. (2022) designed a specialization-generalization training strategy, called match prompt, to handle question-answering tasks. Shi et al. (2022b) presented a soft prompt-based joint learning method for cross-domain aspect term extraction. By incorporating external linguistic features, the proposed method could learn domain-invariant representations between source and target domains through multiple objectives. In view that previous prompt learning methods often used text information, Yang et al. (2023) introduced a multi-modal probabilistic fusion prompt learning approach, which provided diverse cues for multi-modal sentiment detection. However, they did not consider the acoustic or context information. Prompt learning-based approaches have achieved state-of-the-art results in sentiment analysis and emotion recognition. Inspired by this, we argue that the recent prompt learning-based models have not fully considered a new research problem, i.e., multi-modal emotional prompt learning. This raises our research question: *can we design a multi-modal prompt learning model for ERC?*

To answer this question, we aim to introduce multi-modal prompt learning into the ERC scenario and explore its potential. We propose a **M**ulti-modal **A**ttentive **P**rompt (MAP) learning framework for few-shot emotion recognition in conversations. The MAP framework comprises four key modules: a multi-modal feature extraction module, a multi-modal prompt generation module, an attention selection module, and an inference module. First, the multi-modal feature extraction module embeds texts, videos, and speeches into the vectors, referred to as text, video, and speech tokens. Second, based on these multi-modal tokens and contextual utterances, the multi-modal prompt generation module designs six contextual multi-modal prompts. Third, an attention mechanism selectively fuses these prompts to produce a weighted prompt. Finally, we feed this multi-modal prompt into PLMs (specifically RoBERTa and obtain the emotion label (Liu et al., 2019)).

We conduct empirical experiments on two benchmark datasets, MELD and IEMOCAP. A wide range of state-of-the-art baselines, including text-CNN (Kim, 2014), speech-LSTM (Graves et al., 2013), multi-modal CNN (Song et al., 2018b), attention-based bidirectional GRU (BiGRU+Att) (Liu et al., 2020), DialogueRNN (Majumder et al., 2019), DialogueGCN (Ghosal et al., 2019), MMGCN (Hu et al., 2021), BERT (Devlin et al., 2019), EmoBERTa, XLM-T (Barbieri et al., 2022) and EmoCaps (Li et al., 2022b), are compared with the proposed MAP model. The experimental results demonstrate the effectiveness of the proposed MAP model, with a margin of 3.5% and 0.4% in terms of micro F1 over the SOTA system. Additionally, we showcase the superiority and potential of prompt learning over fine-tuning framework through a series of sub-experiments.

1.1 Our Contributions

The main innovations of this work are as follows:

- We make the first attempt to simultaneously incorporate the contextual dependency and multi-modal interaction into a multi-modal prompt learning model.
- We treat ERC as a cloze problem and design six multi-modal templates.
- We validate the effectiveness of the proposed model by applying it to ERC. Experimental results demonstrate that the proposed model outperforms state-of-the-art baselines.

We clarify our innovation from three perspectives in detail.

(1) The difference from other studies. The recent multi-modal emotion recognition approaches have heavily dependent on the availability of sufficient labeled samples. Meanwhile the recent prompt learning approaches have focused on text sentiment analysis, where how to design the multi-modal emotional prompts still needs to be explored. To this end, we propose a novel multi-modal attentive prompt learning model to incorporate the conversational context, multi-modal information and multi-prompts into a unified framework. We argue that our proposed model is quite different from the existing models by analyzing and discussing the related studies (Sec. 2).

(2) The novel research question. The existing multi-modal emotion recognition approaches suffer from one major issue: the large scale datasets are always needed. However, due to the intricacy and subjectivity of emotional understanding, creating and annotating a large-scale, high-quality ERC dataset is a challenging endeavor with substantial annotation and time costs. This leaves us one novel question: How to deal with few-shot emotion recognition in conversations? To answer this question, we make the first attempt to proposed an attention based multi-modal prompt learning model. This focus on few-shot learning is a key differentiator from previous works, as it enables effective emotion recognition even with limited labeled data.

(3) The novel proposed model. Our work introduces a novel approach for multi-modal prompt generation, creating six manually-designed prompts that capture the diverse aspects of conversational emotion. This multi-modal prompt generation module is an innovative contribution, as it considers multiple modalities, such as text, visual, and acoustic inputs, to provide comprehensive cues for emotion recognition. This approach enhances the model's

ability to capture nuanced emotional information in conversations. In addition, the inclusion of an attention mechanism for prompt aggregation is another novel aspect of our work. This mechanism allows the model to focus on the most relevant prompts during the emotion inference process, improving the model's performance and interpretability. We offer extensive experiments to prove that it achieves the best performance across two datasets, and has the fewest parameters and shortest training time. This also proves the effectiveness of the proposed MAP model.

The rest of this paper is organized as follows. Section 2 briefly outlines the related work. Section 3 describes the proposed multi-modal attentive prompt (MAP) learning framework in detail. In Section 4, we report the empirical experiments and analyze the results. Section 5 concludes the paper and points out future research directions.

2. Related Work

We review related studies on multi-modal emotion recognition (including ERC) and prompt learning.

2.1 Early Emotion Recognition

Early studies in emotion recognition primarily focused on uni-modal content, such as text, images, and speech (Mohammad & Turney, 2013; Saxena et al., 2020). Traditional text-based emotion recognition methods relied on words, phrases, and their associated semantics (Mohammad & Turney, 2010; Mohammad & Kiritchenko, 2015; Mohammad et al., 2015). In the case of image-based emotion recognition, early approaches used handcrafted visual features like SIFT, Gabor, LBP, and coupled them with traditional classifiers such as SVM and Random Forests (Mohammad et al., 2018; Zhao et al., 2014). Speech emotion recognition methods of the past extracted acoustic cues like pitch, intensity, tempo, spectral features, and modeled them using techniques like Gaussian mixture models and support vector machines (Mohammad, 2016).

However, this limited unimodal perspective falls short in capturing the nuances of complex human sentiment. The utilization of multiple modalities offers a more comprehensive avenue for conveying rich emotional information, providing vivid descriptions, and uncovering hidden insights that text alone may not capture. Multimodal data's interplay between complementarity and redundancy allows for the modeling of dynamic correlations and interactions across different modalities. Recent advancements in deep multimodal fusion have paved the way for integrated modeling of linguistic, visual, and acoustic cues, promising enhanced capabilities in understanding emotions.

2.2 Multi-Modal Emotion Recognition

Multi-modal emotion recognition aims to determine human basic emotions, such as sadness, surprise, happiness, by analyzing multifarious source samples, such as psychological signals, multi-modal documents, and dialogues (Ma et al., 2023; Liu et al., 2023). It is often viewed as a fine-grained classification task. In earlier studies, machine learning based approaches occupied the mainstream paradigm and achieve remarkable success. For instance, Chuang & Wu (2004) constructed a multi-modal emotion recognition framework based on speech signals and textual documents. Rozgic et al. (2012) incorporated binary SVMs into decision trees as tree nodes to address multi-class emotion recognition problems. Traditional machine learning relies heavily on feature engineering, with hand-crafted features prone to missing emotional cues or overfitting nuances of the training set. However, machine learning models can be more interpretable, lightweight, and data-efficient than large deep neural networks, enabling faster iteration and deployment with less dependence on massive labeled datasets.

Recently, with the advent of deep learning, CNN, RNN, and their multifarious variants have been widely used to extract multi-modal features and build multi-modal emotion recognition framework (Li et al., 2020, 2021, 2022a; Liu et al., 2021; Zhang et al., 2021a; Liang et al., 2023). A few representative works are: Fan et al. (2016) adopted CNN to extract appearance and motion features, and fed them into the RNN for capturing sequence features for video emotion recognition. Kollias & Zafeiriou (2020) utilized a similar manner, training a deep CNN to extract low- and mid-level features, and using RNN subnets to make emotion predictions. Zhang et al. (2021a) proposed a multi-task learning framework that leveraged soft attention and multi-head self-attention to solve depression and emotion detection. Though deep neural networks can achieve high accuracy for emotion recognition, they lack transparency and interpretability compared to machine learning, with complex black-box models making it hard to diagnose failures or biases. However, deep learning automatically extracts optimal features without extensive feature engineering, and can effectively model subtle emotional cues and nuances when trained on large diverse datasets. In summary, deep learning brings performance benefits but sacrifices interpretability, heavily relying on big data rather than human-crafted heuristics to learn representations.

Now, the era of emotion recognition has moved to emotion recognition in conversations, due to the booming of pre-trained language models and social interaction. Emotion recognition in conversation (ERC) has become a popular research topic. First, a few conversational datasets were proposed to support the development of ERC. The representative datasets contain MELD, IEMOCAP, etc. For example, Jia et al. (2022) constructed a multi-modal emotion and desire recognition dataset, called MSED. It consisted of 9,190 text-image pairs collected from a wide range of social media resources, e.g., Twitter, Getty Image, Flickr. Specially, they set a list of keywords with strongly desire expression based on 16 basic desires theory, e.g., $curiosity$, $romance$, $family$, $vengeance$, etc.

Based on the conversational datasets, an increasing number of deep learning based ERC approaches have been proposed. They used and designed different deep learning models to deal with the core problems. Such approaches can be categorized into three main paradigms: RNN-based, Transformer-based, and Graph Learning-based models. RNN based approaches aim to capture the sequential features across utterances. For example, Majumder et al. (2019) described a DialogueRNN model that kept track of the individual party states throughout the conversation and used this information for ERC. Zhang et al. (2019, 2021c) designed a quantum-inspired interactive network (QIN) model for conversational emotion recognition and showed its effectiveness. In addition, Transformer based approaches aim to leverage the advance of Transformer architecture to model the contextual feature. For example, Wang et al. (2022a) proposed to mitigate multi-bias knowledge from Transformer for emotion recognition in conversations. They proposed a series of approaches to mitigate five typical kinds of bias in textual utterances (i.e., gender, age, race, religion and LGBTQ+) and visual representations (i.e, gender and age). Similarly, Ma et al. (2022) also proposed a multiview network (MVN) to explore the emotional representation of a query from two different views, i.e., word- and utterance-level views.

With the development of graph learning, an increasing number of approaches aim to treat words/utterances as nodes, treat their relations as edges, and propose various graph neural networks to solve this ERC problem. For example, Zhang et al. (2021b) also designed the first quantum-inspired multi-task learning framework for sarcasm detection and emotion recognition in conversations. Ishiwatari et al. (2020) presented a relational position encodings-based graph attention network (RGAT), which could save sequential information reflecting the relational graph structure. Ghosal et al. (2019) treated each utterance as a vertex and constructed a dialogue graph. Then, they fed this graph to a graph convolution network and achieved state-of-the-art performance. Lu et al. (2020) presented an iterative emotion interaction network, which explicitly modeled the emotion interaction between utterances. Tu et al. (2022) proposed a context- and sentiment-aware graph attention network, which attempted to capture the sentimental consistency and context information.

2.3 Prompt Learning

Prompt learning-based approaches have emerged as the mainstream paradigm, yielding satisfactory performance in many NLP tasks. There are two kinds of prompt learning approaches: hard and soft prompt learning. Hard prompt learning approaches aim to manually design various prompts to make the downstream task similar to language modeling. For instance, Xu et al. (2022) designed a specialization-generalization training strategy, match prompt, to solve the question-answering task. Yi et al. (2022) developed a contextual information and commonsense-based prompt learning model for conversational sentiment analysis, demonstrating superior performance over state-of-the-art models. Deng et al. (2022b) also introduced a prompt tuning method that mimicked the pre-training objective of contrastive language-image pre-training (CLIP). It thus could leverage the rich image and text semantics for image emotion classification. Wang et al. (2022b) constructed a dynamic virtual template with label words and developed a hierarchy-aware prompt tuning method to handle text classification tasks.

In contrast, soft prompt learning refers to a methodology that involves the use of trainable vectors or representations to guide the generation of prompts in natural language processing tasks. Instead of using fixed or hard-coded prompts, soft prompt learning allows the model to learn the optimal prompts by updating the parameters during training. In soft prompt learning, the prompts are typically represented as continuous vectors that are optimized alongside other model parameters. These vectors capture the contextual information and provide guidance to the model for generating meaningful and contextually relevant responses. By adjusting the prompt vectors through training, the model can effectively adapt and generate appropriate responses based on the given input or task. For example, Gu et al. (2022) proposed to pre-train prompts by adding soft prompts into the pre-training stage to obtain a better initialization. Wu & Shi (2022a) presented an adversarial soft prompt tuning method (AdSPT) to better model cross-domain text sentiment analysis and get new state-of-the-art results. Huang et al. (2022) aimed to provide a better initialization, to improve the performance of prompt learning by considering latent structure within the pre-training data. Shi et al. (2022b) presented a soft prompt-based joint learning method for cross-domain aspect term extraction. Specifically, by incorporating external linguistic features, the proposed method could learn domain-invariant representations between source and target domains via multiple objectives. Shin et al. (2020) proposed an automated method to create prompts for a diverse set of tasks based on a gradient-guided search. They showed that their prompts elicited more accurate factual knowledge from PLMs than the manually created prompts. Deng et al. (2022a) presented a prompt-based fine-tuning strategy to learn task-specific sentiment representations while preserving knowledge contained in CLIP, resulting in a conceptually simple but empirically powerful framework for supervised image emotion classification. Zhou et al. (2023) designed two consistency training strategies for prompt learning and conducted experiments on two multi-label emotion classification datasets. The prompting method has been shown to make the language models more purposeful in prediction by filling the cloze or prefix prompts defined.

In summary, the two aforementioned types of studies have made good progress in many NLP tasks and inspired our work. However, to the best of our knowledge, there is a lack of mechanisms to build a multimodal prompt learning framework. Distinct from existing works, we make the first attempt to simultaneously incorporate contextual dependency and multi-modal interaction into a joint prompt learning framework.

3. The Proposed Approach

In this section, we depict the architecture of the proposed MAP model, which leverages textual, visual, and acoustic information.

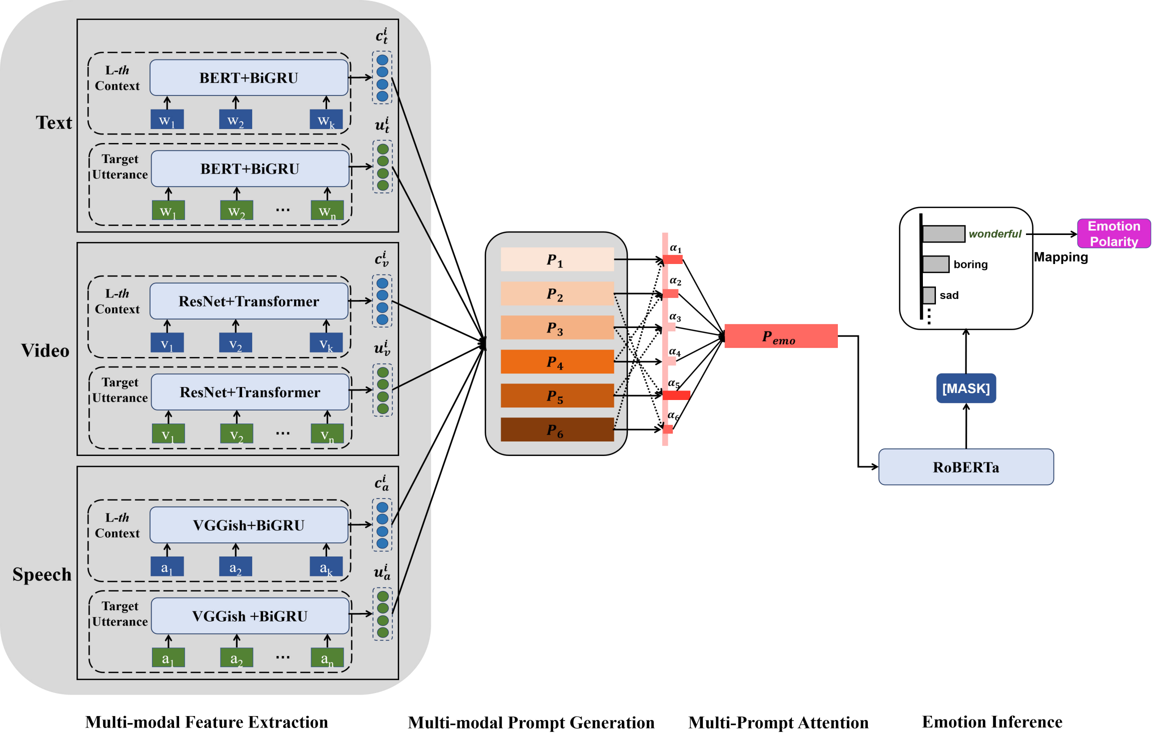

Figure 2: The overall architecture of the MAP model.

3.1 Problem Setup

Assume that a conversational emotion dataset has N multi-modal dialogues, the i^{th} dialogue X^i could be represented as $\left\{X^i = \left(C^i, U^i\right), Y^i\right\}$, where C^i, U^i, Y^i represent the contextual utterance, the target utterance and its label respectively, and $i \in [1, 2, ..., N]$. Both the context and target utterance consist of textual, visual and acoustic modalities, i.e., $C^i = \left(C_t^i, C_v^i, C_a^i\right)$, $U^i = \left(U_t^i, U_v^i, U_a^i\right)$. Given a multi-modal dialogue (including the context C^i and target utterance U^i), how to determine the emotion label of the target utterance Y^i. The research problem can be defined as:

$$\zeta = \prod_i p\left(Y^i | C^i, U^i, \Theta\right) \tag{1}$$

where Θ represents the parameter set.

3.2 Network Description

The network of the MAP model is depicted in Fig. 2. Specially, the MAP framework consists of four building blocks: multi-modal feature extraction module, multi-modal prompt generation module, attention selection mechanism, and inference module. (1) The textual utterance, video clip, and speech segment of the target utterance U_m^i and its context C_m^i are fed into their corresponding encoders to obtain their sequential embeddings, denoted as u_m^i and c_m^i, respectively, where $m \in \{t, a, v\}$. (2) Six multi-modal prompts $\mathcal{P} = \{P_1, P_2, P_3, P_4, P_5, P_6\}$ are manually designed. For example, we may put the video information in the front of the text information to build the first multi-modal prompt, and we can also place the speech information in the front of the text information to build the second multi-modal prompt. The different positions will produce different contextual effects, and thus provide different contextual knowledge for emotion recognition. (3) After initializing such prompts, an attention selection mechanism is applied to generate the attention weights and obtain the weighted prompt $\mathcal{P}_{emo} = Attention(P_1, P_2, P_3, P_4, P_5, P_6)$. (4) According to the prompt, we feed it into a PLM (i.e., RoBERTa, EmoBERTa, etc.), and the PLM predicts the masked tokens based on their contextual tokens. We obtain the emotion label based on the mapping function $F(\cdot)$. We will detail each component in the following subsections. Note that we chose RoBERTa and EmoBERTa as our PLMs among the many BERT-like models because its structure is not only relatively simple, but it can also handle affective information.

3.3 Multi-Modal Feature Extraction

We obtain multi-modal representations of the target utterance and its context in this section.

3.3.1 Textual Utterance

For the target utterance (denoted as $TTar$), suppose there are n terms (words) in the i^{th} target utterance, i.e., $U_t^i = \{w_1, w_2, ..., w_n\}$. Each term $w \in \mathcal{R}^{d_t}$ will be initialized using the pre-trained BERT vectors (Devlin et al., 2019). Then, the word vectors are forwarded to a bi-directional Gated Recurrent Unit (BiGRU) to capture the contexts and get the refined representation of textual utterance:

$$u_t^i = BiGRU(BERT([w_1, w_2, ..., w_n])) \tag{2}$$

Here, we clarify the motivations are: (1) Although BERT can provide contextual representations of individual sentences, it may not fully capture the interplay and dependencies between words in the different contexts, since it was pre-trained using other datasets. By using a GRU, we can fine-tune a sequential model to incorporate the previous context and the target utterance in a dynamic and adaptive manner. We take a “pre-trained+fine-tune” way to obtain the utterance representation without re-training all the parameters of BERT. (2) BERT and GRU have different strengths and capabilities. By combining the strengths of both models, we can potentially achieve a more powerful representation.

To model the contextual information (denoted as $TCon$), we take the contextual utterances appearing in a fixed-window of length L into account. We will tune the length L to find the optimal contextual information, i.e., $\mathcal{C} = \left\{C_t^1, C_t^2, ..., C_t^L\right\}$. Assume that the i^{th} context is composed of k words, i.e., $C_t^i = \{w_1, w_2, ..., w_k\}$. We also use BERT to obtain the pre-trained word embeddings, i.e., $H_{c_i} = [h_t^{w_1}, h_t^{w_2}, ..., h_t^{w_k}]$, and thus feed them to the Bi-GRU for learning the feature representation of each contextual utterance, i.e., $[c_t^1, c_t^2, ..., c_t^L]$, which can be formulated as:

$$c_t^i = BiGRU(BERT(C_t^i)) \tag{3}$$

Note that the default window length is set to $L = 1$. But we will try different window lengths in the experiments, i.e., Sec. 4.6.

3.3.2 Visual Utterance

For the i^{th} video in the dataset (denoted as $VTar$), we assume that it contains n clips, i.e., $U_i^v = \{v_1, v_2, ..., v_n\}$. To ensure the effectiveness of visual representation, we choose to employ the pre-trained visual language model (ResNet50 as default in this work) to extract features from each video clip and produce the corresponding features. Then, the clip vectors are forwarded to a Transformer encoder, composed of six stacked encoder layers, to capture the long-range context within the video. Each layer in the encoder consists of a multi-head self-attention module (MSA) and a multilayer perceptron (MLP). The output can be expressed as:

$$u_v^i = Transformer(ResNet(\{v_1, v_2, ..., v_n\})) \tag{4}$$

In order to preserve the positional information, the visual feature vector u_v^i is added by a learnable positional embedding u_p that is randomly initialized, i.e., $u_v^i = u_v^i \oplus u_p$. For the contexts (denoted as $VCon$), we adopt a similar manner approach to obtain the vector representation, c_v^i.

3.3.3 Acoustic Utterance

For the i^{th} acoustic counterpart (denoted as $ATar$), we assume that it contains n frames, i.e., $U_i^a = \{a_1, a_2, ..., a_n\}$. To ensure the effectiveness of acoustic representation, we choose to employ the VGGish network pre-trained on Audio Set, which is a large-scale labeled audio dataset released by Google in 2017, to extract vocal features. The network outputs a 128-dimensional feature vector for each temporal frame. To capture the contextual knowledge, we feed the features through a BiGRU network to obtain the acoustic representation of the target speech:

$$u_a^i = BiGRU(VGGish([a_1, a_2, ..., a_n])) \tag{5}$$

To model the contextual acoustic information (denoted as $ACon$), we choose to use the same pre-trained model and the BiGRU network (refer to Eq.5) to obtain the context representation, c_a^i.

3.4 Prompt Generation

In prompt engineering, the input sentences are formalized as the natural language template, and the emotion recognition task is treated a cloze task. The template provides a background description of the current task, and label words are the high-probability words predicted by PLMs in the given context. emotion labels are linked to the label words through a mapping function. In this work, we denote the target multi-modal utterance as U_m^i, its contexts as C_m^i and the output label as $Y^i \in \{anger, disgust, fear, joy, neutral, sadness, surprise\}$. We suggest that an effective approach is to place the contextual utterances into the former parts of the prompt template and put the target utterance into the latter parts. In this manner, we can bridge the gap between the task of ERC and a cloze task and build the mapping between them. Consequently, one can use the pre-trained knowledge from PLMs to obtain the emotion labels of the target utterance by filling in the blank.

The procedure of multi-modal prompt generation is detailed here. Given a PLM $\mathcal{M}$, i.e., RoBERTa, and its vocabulary $\mathcal{V}$, we design six prompts that consist of different template functions $T(\cdot)$ to convert the target utterance U^i and its context to six prompt inputs $\mathcal{P} = T(U_m^i, C_m^i)$ with the [MASK] token. In order to construct a multi-modal prompt, our approach is to leverage visual and acoustic knowledge as the auxiliary information. Since the textual, visual, and acoustic features serve the same utterance, one natural choice is to combine them and build a multi-modal prompt. Additionally, considering the importance of contextual information, we incorporate contextual utterances into the multi-modal prompt. We intuitively construct six prompts as the default setting. We have tried different numbers of prompts in the experiments (please refer to Sec. 4.8).

More specifically, these are:

$$\begin{aligned}
P_1:\ &\{Speech:[ACon][ATar].\ Video:[VCon][VTar].\ Text:[TCon][TTar].\\
&Overall, the\ emotion\ is\ [MASK].\}\\
P_2:\ &\{Video:[VCon][VTar].\ Speech:[ACon][ATar].\ Text:[TCon][TTar].\\
&Overall, the\ emotion\ is\ [MASK].\}\\
P_3:\ &\{Video:[VCon][VTar].\ Speech:[ACon][ATar].\ Text:[TCon][TTar].\\
&Which\ emotion\ is\ expressed?\ [MASK].\}\\
P_4:\ &\{Speech:[ACon][ATar].\ Video:[VCon][VTar].\ Text:[TCon][TTar].\\
&Which\ emotion\ is\ expressed?\ [MASK].\}\\
P_5:\ &\{Speech:[ATar].\ Video:[VTar].\ Text:[TCon][TTar].\\
&Which\ emotion\ is\ expressed?[MASK].\}\\
P_6:\ &\{Speech:[ATar].\ Video:[VTar].\ Text:[TCon][TTar].\\
&Overall, the\ emotion\ is\ [MASK].\}
\end{aligned}$$

The actual prompt inputs consist of the embeddings of {*textual, visual, acoustic*} target utterance, the embeddings of {*textual, visual, acoustic*} contextual utterances, the embeddings of the [MASK] token and the embeddings of two positional tokens [CLS] and [SEP]. Therefore, the actual prompt inputs can be represented as: $[e\left([CLS]\right), e\left(\{P_1, P_2, ..., P_6\}\right), e\left([MASK]\right)$ $,e\left([SEP]\right)]$, where $e\left(\cdot\right)$ represents the embedding functions. Now, we have obtained six multi-modal prompts, which will be used to calculate the final weighted prompt.

We clarify that the different positions create distinct contextual effects, providing varied contextual knowledge for emotion recognition. Here are some explanations about why all possible orders and combinations are not taken into account: (a) Feasibility and resource limitations in experimental design: In the context of multi-modal prompt learning, considering factors such as data availability and computational resources, it is not feasible to consider all possible orders and combinations. Therefore, in our experimental design, we selected a representative set of variations that cover different orders and combinations of speech, video, and text information. This allows us to conduct experiments within practical constraints while still obtaining meaningful and interpretable results. (b) By selecting variations with different orders, we facilitate the model's learning of interactions and dependencies between different modalities in the multi-modal input. This diversity aids in improving the model's generalization performance and its ability to learn richer feature representations. However, due to practical considerations and experimental needs, it is not possible to consider all possible orders and combinations.

3.5 Multi-Prompt Attention

The aforementioned six prompts are designed based on manual effort and guesswork, potentially limiting their effectiveness due to unsuitable prompts. To address this issue, we introduce a self-attention mechanism to control the influence of the multi-modal prompts set and produce a new fused prompt by calculating the attention scores from the six prompts.

Specifically, we randomly select a prompt from the set of the prompts $\{P_1, P_2, ..., P_6\}$ as the initial prompt $\mathcal{P}_{emo}$, and set it to be learnable. Then, we regard it as $Query$, i.e., $Q_k = W_q\mathcal{P}_{emo}$ and treat the set of six hard prompts as $Keys$ and $Values$, i.e., $K_j = W_kP_j$, $V_j = W_vP_j$, where $j \in [1, 2, ..., 6]$. This results in:

$$\begin{aligned} \alpha_j &= softmax\left(\frac{Q_kK_j}{\sqrt{d_k}}\right)V_j \\ &= softmax\left(\frac{W_q\mathcal{P}_{emo} \cdot W_kP_j}{\sqrt{d_k}}\right)W_vP_j \\ \mathcal{P}_{emo} &= \sum_{j=1}^{6} \alpha_jP_j \end{aligned} \tag{6}$$

where W_q, W_k and W_v are weights. The attention-based prompt is fed through subsequent layers of the PLM $\mathcal{M}$. The PLM $\mathcal{M}$ implicitly models the inter-modal interaction and generates the prediction for the [MASK] token. The inference process is detailed in the following section.

3.6 Emotion Inference

In the inference process, we utilize a PLM (e.g., RoBERTa, EmoBERTa, etc.) and a mapping function (also known as a verbalizer, denoted as F) to predict the emotion of the utterance. The PLM predicts the answer that is most likely to appear at the position of [MASK], based on the embeddings of the input prompt $\mathcal{P}_{emo}$. The mapping function is used to map the answer to the emotion labels (i.e., anger, disgust, fear, joy, neutral, sadness, surprise). It bridges the gap between the conversational input space and the prompt input space. The mapping approach in our work involved manually designing answers for each label based on synonyms. While the selection of synonyms is somewhat intuitive, which are collected from the Oxford English Dictionary or Thesaurus.com[1] which is world's largest online thesaurus. For example, we collect about 55 synonyms for "happy" emotion, 40 synonyms for "sadness" emotion, 45 synonyms for "fear" emotion, and so on. The collecting process is that we take the emotion label word as the input, query it on the website and collect the retrieval results. The synonyms list can be see Appendix.

For example, we set V(sad|unhappy| heartbroken| sorrow) = sadness, V(cheerful|delight|joyful) = joy. Given the answer l, the mapping function outputs the corresponding emotion label. Let the vocabulary of PLM to be $\mathcal{V}$, and Y^i be the emotion label. The inference process can be expressed as:

$$P\left(Y^i|U^i, C^i, \theta, \mathcal{P}_{emo}\right) = \frac{e^{\mathcal{M}(F(l)|\mathcal{P}_{emo})}}{\sum_{l'} e^{\mathcal{M}\left(F(l')|\mathcal{P}_{emo}\right)}} \tag{7}$$

[1] https://www.thesaurus.com/

Where $\mathcal{M}\left(F(l)|\mathcal{P}_{emo}\right)$ represents the predicted probability of the label word l appearing at the position of [MASK]. $F(l)$ maps the predicted word l into one emotion label Y^i. With this, the emotion prediction is obtained.

3.7 Model Training

We use cross entropy with L2 regularization as the loss functions for training and minimize it with different weights.

$$J = -\frac{1}{N}\sum_{i}\sum_{m} Y_i^m log\hat{Y}_i^m + \lambda_r \left\|\theta\right\|^2 \tag{8}$$

where Y_i denotes the ground truth and $\hat{Y}_i$ is the predicted emotion distribution. i is the utterance index, and m is the class index. λ_r is the coefficient for $L2$ regularization. We use Adam to compute the gradients and update all the parameters. To avoid overfitting, we employ a dropout strategy.

4. Experiments

In this section, we conduct extensive experiments to evaluate the performance of the proposed MAP model.

4.1 Experimental Settings

Datasets. MELD[2] (Poria et al., 2019) and IEMOCAP[3] datasets (Busso et al., 2008) are selected as our datasets. MELD contains 13,708 utterances from 1433 dialogues of *Friends* TV series. It annotates each utterance with one of seven emotions (anger, disgust, fear, joy, neutral, sadness or surprise). It contains a total of approximately 33 hours of dialogues.

IEMOCAP is a multi-modal database of ten speakers involved in two-way dyadic conversations. Each utterance is annotated using one of the following emotion categories: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, or others. It contains approximately 12 hours of audio-visual recordings from 5 mixed gender pairs of actors. Each conversation was about 5 minutes long. There are 8425 utterances in total.

MELD and IEMOCAP are multi-modal ERC datasets that involve all the textual, visual, and acoustic information. The dataset details are provided in Table 1.

[2] https://affective-meld.github.io/

[3] http://sail.usc.edu/iemocap/

Dataset	Emotion Category	No. of Utterances		
		Train	Dev	Test
MELD	anger	1109	153	345
	disgust	271	22	68
	fear	268	40	50
	joy	1743	163	402
	neutral	4710	470	1256
	sadness	683	111	208
	surprise	1205	150	281
IEMOCAP	anger	804	#	158
	happiness	377	#	127
	sadness	592	#	191
	neutral	1124	#	357
	fear	572	#	129
	surprise	542	#	104
	other	2486	#	859

Table 1: Training, validation, and test data distribution in the datasets.

Evaluation metrics. Considering the imbalanced sample problem, we choose the **precision**, **recall**, micro **F1**, and balanced **accuracy** as the evaluation metrics.

We clarify the reason to choose micro F1 as the evaluation metrics in our experiment is that micro F1 is more suitable to deal with the imbalanced datasets. Micro F1 and Macro F1 are two common metrics, and often applied to multi-class classification problem. Here are their explanations from sklearn package (which is the most useful and robust library for machine learning in Python). Micro F1 calculates F1 globally by counting the total true positives, false negatives and false positives. Macro F1 calculates F1 for each label, and finds their unweighted mean. Macro F1 does not take label imbalance into account. So micro F1 reflects the accuracy on imbalanced data better than macro F1. Based on this, the micro-F1 will adequately capture this class imbalance and take the imbalance problem into consideration. In our experiment, both datasets, i.e., MELD and IEMOCAP, are imbalanced datasets, where the number of each class is not equal. In addition, the balanced accuracy is the average between the sensitivity and the specificity, which measures the average accuracy obtained from both the minority and majority classes. Choosing balanced accuracy as an evaluation metric is particularly beneficial when dealing with imbalanced datasets, as it ensures a more fair and comprehensive assessment of the model's performance across all classes. In addition, all baselines are evaluated using the same metric, so the performance gap among them is fair.

Hyperparameter Setting. The textual, visual, and acoustic inputs are initialized with BERT, ResNet, and VGGish, respectively. All weight matrices are given their initial values by sampling from a uniform distribution $U(-0.1, 0.1)$. The optimal learning rate is set to 4e-7 for MELD dataset and 6e-7 for the IEMOCAP datasets. The batch size is set to 1 and the number of epochs is set to 50 for MELD and 150 for IEMOCAP. The dropout rate is set to 0.1 for MELD and 0.5 for IEMOCAP.

4.2 Baselines

We present and compare our model with a number of baselines. They are listed as follows.

(1) **Text-CNN:** we apply a deep convolutional neural network (CNN) on each utterance to extract the textual features and put them through the softmax classifier to get the decision.

(2) **Speech-LSTM:** we forward the acoustic counterpart represented by VGGish, into an LSTM to obtain the emotion classification.

(3) **Multi-modal CNN (Zhang et al., 2020):** we adopt two deep CNNs to extract textual and visual features and use the LSTM to extract acoustic features. Then, we merge them and feed the multi-modal features into the softmax function for emotion recognition.

(4) **BiGRU+Att:** we use two separate bidirectional GRUs to extract the textual and acoustic features and a CNN-BiGRU to extract video features. It forwards the concatenated multi-modal representation through a softmax function for emotion detection.

(5) **SVM+BERT (Zhang et al., 2022):** we use BERT to produce the utterance embeddings and use an SVM classifier to obtain the emotion prediction.

(6) **DialogueRNN (Majumder et al., 2019):** we implement a famous sequence-based ERC model, which uses three GRUs to model the speaker, the context, and the emotion behind the preceding utterances.

(7) **DialogueGCN (Ghosal et al., 2019):** we implement a state-of-the-art graph learning-based ERC model, which models the conversation using two-layer graph neural networks.

(8) **XLNet (Yang et al., 2019):** the XLNet baseline with the original segment recurrence and vanilla self-attention, initialized with the pre-trained parameters of the XLNet-base.

(9) **MMGCN (Hu et al., 2021):** we report a state-of-the-art baseline. MMGCN incorporates textual and visual knowledge into a unified ERC model.

Three open source emotion recognition baselines are listed below:

MELD dataset	**Model**	**Evaluation metric**			
		Precision	Recall	F1	Accuracy
Emotions	Text-CNN	0.520	0.548	0.532	0.548
	Speech-LSTM	0.481	0.503	0.493	0.497
	Multi-modal CNN	0.536	0.547	0.543	0.547
	BiGRU+Att	0.556	0.537	0.536	0.541
	SVM+BERT	0.611	0.629	0.622	0.619
	DialogueRNN	0.559	0.581	0.570	0.576
	DialogueGCN	0.580	0.584	0.581	0.590
	XLNet	0.609	0.617	0.617	0.620
	MMGCN	-	-	0.587	-
	EmoCaps	-	-	0.640	-
	XLM-T	0.601	0.634	0.615	0.633
	EmoBERTa	0.761	0.757	0.757	0.757
	Text-MAP	0.600	0.628	0.628	0.628
	Text&Video	0.602	0.624	0.624	0.624
	Text&Speech	0.590	0.615	0.615	0.615
	(RoBERTa) MAP	0.602	**0.629**	0.629	**0.629**
	(EmoBERTa) MAP	**0.784**	**0.785**	**0.784**	**0.785**

IEMOCAP dataset	**Model**	**Evaluation metric**			
		Precision	Recall	F1	Accuracy
Emotions	Text-CNN	0.534	0.564	0.538	0.564
	Speech-LSTM	0.421	0.533	0.448	0.533
	Multi-modal CNN	0.518	0.546	0.521	0.546
	BiGRU+Att	0.550	0.547	0.546	0.554
	SVM+BERT	0.657	0.649	0.641	0.649
	DialogueRNN	0.619	0.631	0.627	0.634
	DialogueGCN	0.640	0.644	0.641	0.653
	XLNet	0.609	0.611	0.613	0.607
	MMGCN	-	-	0.662	-
	EmoCaps	-	-	0.718	-
	XLM-T	0.546	0.546	0.545	0.546
	EmoBERTa	0.658	0.666	0.667	0.666
	Text-MAP	0.611	0.594	0.594	0.594
	Text&Video	0.588	0.590	0.590	0.590
	Text&Speech	0.416	0.492	0.492	0.492
	(RoBERTa) MAP	0.615	0.631	0.631	0.631
	(EmoBERTa) MAP	**0.715**	**0.721**	**0.721**	**0.719**

Table 2: Performance of all the baselines on the MELD and IEMOCAP datasets. The best-performing system is indicated in bold.

(10) **EmoCaps (Li et al., 2022b):** we report a state-of-the-art baseline. EmoCaps extracts emotion vectors through the Emoformer structure and obtains the emotion classification results from a context analysis model.

(11) **EmoBERTa** [4]: it a simple yet expressive scheme of solving the ERC (emotion recognition in conversation) task. By simply prepending speaker names to utterances and inserting separation tokens between the utterances in a dialogue, it can learn intra- and inter- speaker states.

(12) **XLM-T (Barbieri et al., 2022)**[5]: it is designed to understand and classify emotions, which supervised fine-tunes RoBERTa on multi-modal sentiment datasets.

One uni-modal and two bi-modal baselines are listed below:

(13) **Text-MAP:** we only design many textual prompts instead of using the multi-modal prompts and making emotional inferences.

(14) **Text&Video:** we add visual information to build the bi-modal prompts and make emotional inferences.

(15) **Text&Speech:** the acoustic information is used to design bi-modal prompts and construct the fused attentive prompt.

4.3 Results on the MELD Dataset

The first set of experiments is conducted on the MELD dataset, which contains more samples than IEMOCAP. The experimental results are shown in Table 2.

From Table 2, we observe that Speech-LSTM performs poorly, indicating that relying solely on acoustic features is insufficient for the emotion label analysis. In contrast, Text-CNN performs better, as expected, because textual information often plays a more critical role in emotion detection. The introduction of the attention mechanism allows BiGRU+Att to outperforms Text-CNN and Speech-LSTM by modeling the context information and assign different weights to contexts. Multi-modal CNN surpasses BiGRU+Att by combining text, video, and speech information, demonstrating the importance of multi-modal fusion in emotion detection. The multi-modal representation provides more complementary knowledge than the uni-modal representation. Two well-known ERC baselines, DialogueRNN and DialogueGCN, show significant progress over the aforementioned baselines. Both consider contextual utterance and the speaker information, make their models more complex and capable of capturing the inter-speaker and intra-speaker dependencies. But neither of them exceeds 60% in terms of F1. MMGCN performs very poor and gets comparable results against DialogueGCN.

Incorporating pre-trained language models into ERC allows SVM+BERT to outperform DialogueGCN by a large margin of nearly 7.0% in terms of micro-F1 score. This

[4] https://huggingface.co/tae898/emoberta-large

[5] https://github.com/cardiffnlp/xlm-t

significant improvement can be attributed to BERT's robust capacity for contextual representation.Similarly, XLNet surpasses DialogueRNN and DialogueGCN as it learns intra- and inter-speaker states and context to predict the current speaker's emotion in a pre-trained language modeling manner. XLM-T achieves the similar performance against XLNet, as it is mainly fine-tuned for multilingual sentiment analysis dataset, which may not adaptable to multi-modal sentiment analysis. EmoCaps performs well due to its effective use of multi-modal dependencies and speaker information. It achieves F1 result of 64.0%, significantly higher than the above-mentioned baselines. It can effectively preserves the sequential order of utterances and enables consecutive utterances to share information. EmoBERTa performs the best, and achieves the best classification performance among all the baselines. Because it can learn intra- and inter- speaker states and context to predict the emotion of a current speaker.

Finally, Text-MAP performs well compared to XLNet and MMGCN. However, Text& Video and Text&Speech show slightly decreases, suggesting our proposed multi-modal prompts need further refinement to better model multi-modal complementary. The proposed RoBERTa based MAP model takes the first step towards building a tri-modal prompt learning model, achieving the third best classification results across all metrics and significantly outperforms all the baselines. Using MAP, F1 increases by 0.6% compared to SVM+BERT. Notably, we do not fine-tune the pre-trained language model. In addition, in view that RoBERTa is not a SOTA emotional PLM, we also propose a EmoBERTa based MAP model, which treats EmoBERTa as the base model. The improved EmoBERTa based MAP model achieves the best classification performance, and defeats all the baselines. Compared with the EmoBERTa model, our MAP model obtains significant improvements of 3.56% and 3.69% in terms of F1 and accuracy scores. Overall, we attribute the main improvements to both the multi-modal prompt and the attention mechanism, ensuring that the MAP model can effectively adjust to downstream tasks. A detailed ablation study is provided in subsection 4.5.

4.4 Results on the IEMOCAP Dataset

Table 2 presents the performance comparison of the MAP model with the baselines on the IEMOCAP dataset, another widely used dyad conversational emotion dataset.

From Table 2, we can first observe the poor performance of Speech-LSTM. Text-CNN works better than Speech-LSTM. Multi-modal CNN can produce improved results over Speech-LSTM but fails to improve the performance over Text-CNN. One possible reason is that the simple feature-level fusion method cannot effectively capture the correlation between multi-modalities. It is necessary to develop an alternative multi-modal fusion approach. BiGRU+Att outperforms Speech-LSTM due to the introduction of an attention mechanism. Unlike their performance on MELD, XLNet, XLM-T and DialogueRNN perform much worse on IEMOCAP, with a sharp drop in classification performance. The reason is that the dark visual document cannot provide adequate information for multi-modal feature fusion. Another possible reason is that IEMOCAP provides fewer training samples, so they cannot find the global optimum to fine-tune their pre-trained language models. DialogueGCN outperforms them where it reaches 64.1% and 65.3% in terms of F1 and accuracy. Furthermore, EmoBERTa and MMGCN perform better than the above models, where EmoBERTa performs the better than MMGCN. The state-of-the-art approach EmoCaps performs quite well, and gain the highest scores. It significantly outperforms all of the above baselines (71.8% *vs* 66.7% of EmoBERTa). This proves the effectiveness of the "fine-tuning" paradigm.

Finally, aiming to establish a multi-modal prompt learning framework, our proposed RoBERTa based MAP model outperforms XLNet, DialogueRNN and XLM-T by 2.9%, 0.6% and 15.7% in terms of the F1 score. However, our proposed RoBERTa based MAP model obtains comparable results against DialogueGCN and SVM+BERT. But our MAP does not need to train the model. The RoBERTa based MAP model is worse than EmoBERTa, MMGCN and EmoCaps. There are two reasons: (1) we do not train the RoBERTa model, and only use its pre-trained knowledge; (2) our designed multimodal prompts are not as effective on IEMOCAP as they are on MELD. Because significant interference occurs in our prompts due to the noise in visual information. When the visual information is full of noise, significant interference occurs in our prompts. To fill this gap, we have adopted the EmoBERTa as the base model to build an EmoBERTa based MAP model. This new MAP model obtains an improvement of 8.1% in terms of F1 score over EmoBERTa. This proves that the effectiveness of multi-modal prompt learning. In addition, the proposed EmoBERTa based MAP model outperforms EmoCaps with a slight improvement of 0.4% (72.1% *vs* 71.8%). It achieves the state-of-the-art results on IEMOCAP dataset. Moreover, it has fewer parameters and needs less training time (see Sec. 4.7). Hence, it can be more suitable for few-shot or real-time emotion recognition. In addition, we can also use stronger PLMs as the base model to further enhance the performance. This demonstrates its potential in ERC.

4.5 Ablation Study

In order to explore the contribution of different components of the proposed MAP model, we design several submodels by removing one component at a time: (1) *No-Atten* which removes the attention mechanism from the MAP model and randomly selects one of the six prompts as the final prompt; (2) *No-Prompt* which removes all of the multi-modal prompts and merges the multi-modal representation; (3) *Uni-modal Prompt* which removes the visual and acoustic information from the MAP model and only keeps the textual prompts.

The experimental results are shown in Table 4. We can observe that the MAP model achieves the best performance on both datasets, indicating that all the components contribute to the classification performance. We then notice that the No-Prompt model performs the worst on both datasets, which proves that the multi-modal prompt has the most significant contribution to the classification performance. The No-Atten and Uni-modal

Dataset	**Model**	**Metric**	
		F1	Accuracy
MELD	EmoBERTa	0.757	0.757
	SVM+EmoBERTa	0.763	0.760
	MAP	0.784	0.785
IEMOCAP	EmoBERTa	0.667	0.666
	SVM+EmoBERTa	0.684	0.685
	MAP	0.721	0.719

Table 3: Comparison between different EmoBERTa based models.

Prompt models achieve the second-worst results, demonstrating that attention and multi-modal information also play essential roles in improving performance. However, No-Atten performs worse than Uni-modal Prompt, suggesting that the attention mechanism contributes more than the multi-modal information. This supports our previous argument that the proposed multi-modal fusion is too naive to model the inter-modal complementarity. So when the visual information is full of noise, the effect of the multi-modal prompt will be small. Overall, the ablation study suggests that a) an effective prompt learning model could help the pre-trained language model better ”understand” multi-modal documents; b) the attention mechanism is helpful in selecting prompts; (c) a refined multi-modal fusion approach is needed.

In addition, we have improved SVM+BERT by relpacing BERT with EmoBERTa. The target is to offer a fair comparison between our MAP+EmoBERTa and other EmoBERTa based models. The experimental results are shown in Table 3. We can notice that the proposed MAP model outperforms the new SVM+EmoBERTa baseline on both datasets with a significant improvement of 2.7% and 4.9% (78.5% *vs* 76.0%; 71.9% *vs* 68.5%). This comparison offers a more holistic view of our model’s capabilities and its potential to outperform baselines by leveraging domain-specific LLMs. Our model is designed as a flexible and versatile plug-and-play plugin framework that can seamlessly integrate with different pre-trained language models. By including EmoBERTa, we illustrate that our MAP method can effectively harness the power of stronger, emotion-specific LLMs, offering improved results with shorter training time. This adaptability and compatibility with various language models are key advantages of our proposed framework.

4.6 The Impact of the Window Length

In MAP, the default window length is set to $L = 1$. To find the optimal contextual information, we have tried different window lengths ranging from $\{1, 2, 3\}$. We incorporate different contexts into the multi-modal prompts according to different window lengths. The experimental results are shown in Table 5. ”L=2” means treating the previous two utterances as the context. ”L=3” means using the past three contexts to learn contextual representation.

Dataset	Model	Metric	
		F1	Accuracy
MELD	No-Atten	0.627	0.627
	No-Prompt	0.607	0.607
	Uni-modal Prompt	0.628	0.628
	MAP	0.629	0.629
IEMOCAP	No-Atten	0.588	0.588
	No-Prompt	0.529	0.529
	Uni-modal Prompt	0.594	0.594
	MAP	0.631	0.631

Table 4: Ablated MAP for both MELD and IEMOCAP datasets.

Dataset	Context Size	Sentiment	
		F1	Accuracy
MELD	L=1	0.629	0.629
	L=2	0.627	0.627
	L=3	0.601	0.601
IEMOCAP	L=1	0.631	0.631
	L=2	0.530	0.530
	L=3	0.509	0.509

Table 5: Effect of context range.

From Table 5, we observe that MAP with one context performs the best for emotion recognition on MELD and IEMOCAP. This indicate that (1) modeling one context is sufficient in prompt learning, and (2) modeling too many historical utterances can introduce noise that may negatively impact performance. MAP with two contexts obtains better results than MAP with three contexts, with improvements of 2.1% and 3.6% in terms of F1 score. MAP with three contexts performs the worst among the three baselines. This implies that it is an excellent choice to consider one contextual utterance in multi-modal prompt learning.

4.7 Parameter Comparison and Complexity

Since prompt learning requires fewer parameters and less parameter updating compared to previous approaches, we present and compute the number of parameters and the running time of the proposed MAP model and another five state-of-the-art baselines. The experimental results are shown in Table 6.

From Table 6, we see that MAP, with only 1.2M parameters, displays clear advantages in training efficiency over the larger models. The largest model, XLNet (380M parameters), unsurprisingly takes the longest to train - 28.5 hours on MELD, and 18.4 on IEMOCAP.

Dataset	Baselines	Parameters	Time (hour)
MELD	XLNet	380M	28.5
	SVM+BERT	110M	6.2
	XLM-T	85M	4.4
	EmoBERTa	116M	6.4
	EmoCaps	15M	3.9
	MAP	1.2M	2.2
IEMOCAP	XLNet	380M	18.4
	SVM+BERT	110M	4.5
	XLM-T	85M	4.0
	EmoBERTa	116M	4.5
	EmoCaps	15M	2.6
	MAP	1.2M	1.4

Table 6: Comparison of training time.

This reflects its immense complexity coming from the multi-layer transformer architecture. However, this high capacity enables XLNet to achieve state-of-the-art performance on many other NLP tasks. SVM+BERT and EmoBERTa, with 110M and 116M parameters respectively, take 4-6 hours to train. Their mid-size transformer architectures balance complexity and training costs. EmoCaps at 15M parameters sees training times of 2.6-3.9 hours on two datasets. This demonstrates the efficiency benefits of capsule network-based models compared to transformers.

Clearly, MAP provides the best combination of training efficiency and architecture simplicity. With 302x fewer parameters than XLNet, it trains over 10x faster on both datasets. Its ultra-low parameter count enables drastically faster and cheaper training, while preserving strong modeling performance as evidenced by competitive results on these datasets. The design philosophy of prompt learning behind MAP demonstrates how future model development can priorities parameter efficiency without sacrificing effectiveness.

4.8 The Impact of the Number of Prompts

In this subsection, we aim to explore the optimal number of prompts. In this work, the default setting is six multi-modal prompts. We will try different numbers of prompts from one to ten, and the experimental results are shown in Figure 3. While it is also possible to consider the number of prompts as a hyperparameter and tune it during the experimental setup, we opted for a fixed number of prompts for two reasons: (a) introducing the number of prompts as a hyperparameter would significantly increase the search space for optimization. With a fixed number of prompts, we can focus on exploring other important factors and variations in the experimental design. (b) Adding the number of prompts as a hyperparameter could introduce additional complexity to the model and the training process. Keeping it fixed simplifies the experimental setup and reduces the risk of overfitting or model instability.

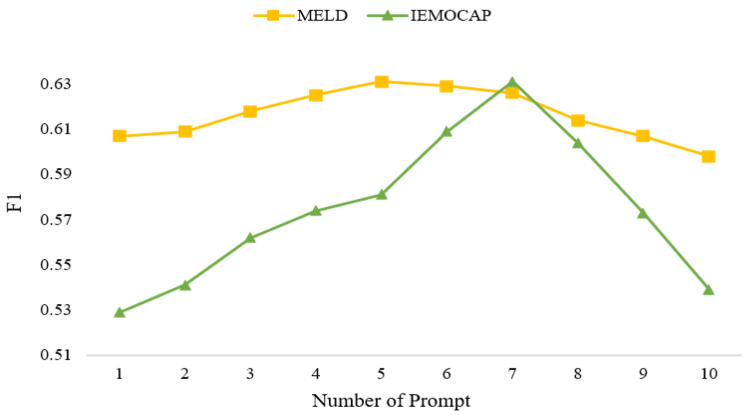

Figure 3: The impact of the number of prompts.

From Figure 3, we observe that the MAP model achieves the best classification performance when the number of prompts is five on MELD and seven on IEMOCAP. When the number of prompts is six or seven on MELD and six or eight on IEMOCAP, the MAP classification performance reaches the second-highest values for both datasets. As the number of prompts increases, the performance of the MAP model exhibits an overall trend of initially rising and then declining. This indicates that having either too many or too few prompts limits MAP from obtaining its maximum advantage. When the number of prompts is too small, the classification task cannot be effectively converted to a cloze task. When there are too many prompt templates, more noise is introduced. When the number of prompt templates reaches six, seven or eight, MAP strikes a balance between introducing knowledge and noise.

4.9 The Effect of other PLMs

In this work, we use the RoBERTa model as the default for predicting emotion labels. Although RoBERTa has apparent advantages in long-range dependency modeling, we are interested in exploring the potential of other strong PLMs. Due to resource and computational limitations, we treated the choice of PLM as a fixed configuration rather than tuning it as a hyperparameter in the paper. There are several reasons for this approach. Firstly, we aim to provide a detailed analysis and comparison specifically for Roberta as the baseline model. Secondly, treating the choice of PLM as a hyperparameter may require larger-scale resources and computations that go beyond our computation scope. As a result, we also experiment with six widely used state-of-the-art PLMs: BERT, Transformer, GPT-2, EmoBERTa, XLM-T, Flan-T5. We re-run the experiments with different PLMs and obtain the experimental results, as shown in Table 7. Here, we clarify when referring to the "standard Transformer", we mean the basic architecture of the Transformer model without any specific modifications or pre-training objectives. We acknowledge that BERT, GPT-2, EmoBERTa and Flan-T5 (Chung et al., 2022) are built on the transformer architec-

Dataset	**PLMs**	**Metric**	
		F1	Accuracy
MELD	Transformer	0.610	0.607
	BERT	0.627	0.629
	GPT-2	0.631	0.633
	EmoBERTa	**0.784**	**0.785**
	XLM-T	0.648	0.648
	Flan-T5	0.612	0.628
	RoBERTa	0.629	0.629
IEMOCAP	Transformer	0.614	0.611
	BERT	0.622	0.619
	GPT-2	0.631	0.633
	EmoBERTa	**0.721**	**0.719**
	XLM-T	0.548	0.545
	Flan-T5	0.471	0.494
	RoBERTa	0.631	0.631

Table 7: The effect of different PLMs.

ture. However, each of these models has specific modifications and pre-training objectives that differentiate them from the basic Transformer.

From Table 7, we can see that the classification performance of Transformer is the worst on MELD since it is the basic pre-trained language model and has the least amount of parameters. BERT outperforms the basic Transformer model but performs worse than RoBERTa. The reason is that RoBERTa is an extension of BERT and can be seen as a refined version of BERT, which was trained on a larger dataset. Hence, it learns more knowledge than BERT. Given that GPT-2 and XLM-T have a significant number of parameters, both achieve better classification performance than the other three models. Despite that Flan-T5 overcomes Transformer, but still performs poor on two datasets. One possible reason is that The Flan-T5 model might have been trained on different types of tasks or datasets, which may not be well-suited for sentiment analysis. Sentiment analysis requires sensitivity to emotions, and datasets from different domains may have varied ways of expressing sentiments, which could negatively impact Flan-T5's performance. In contrast, EmoBERTa based MAP achieves the best performance, and significantly outperforms other PLMs and MMGCN, XLNet, EmoCaps. In addition, it also achieves the best performance on IEMOCAP datasets, and achieves the second best scores among all the baselines in Table 2. This shows the effectiveness of the propsoed multi-modal prompt learning and also proves that our proposed MAP has great potential and can gain better performance by using a stronger PLM.

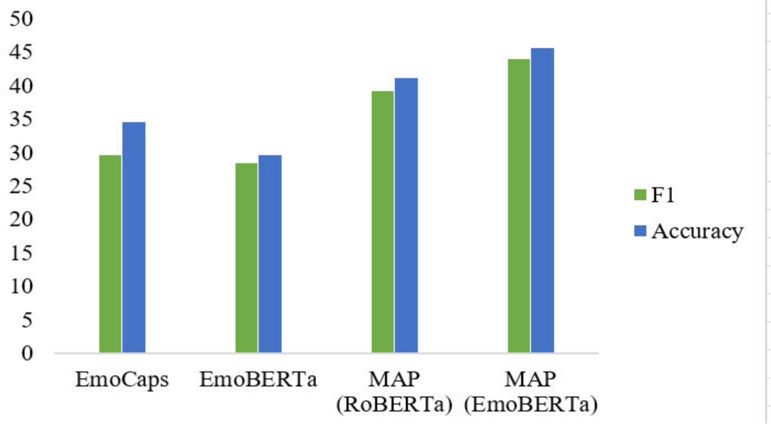

Figure 4: The performance of few-show learning.

4.10 Few Shot Learning

In order to prove the few-shot learning ability of our proposed MAP model, we randomly collect 20 samples from the dataset, where 13 samples are treated as the training set, the remaining 7 samples are treated as the testing set. We re-run the EmoBERTa, EmoCaps model and two MAP models. More specifically, we will re-train EmoBERTa and EmoCaps on this small dataset. As for the MAP models, we can obtain the results directly by taking the prompt as the input. The results are shown in Figure 4. We can notice that EmoCaps and EmoBERTa drops very fast, where their F1 scores are 0.296 and 0.287. The reason is that the EmoCaps and EmoBERTa models overfit on this small dataset. However, our proposed two MAP models outperform them by a large margin, where both of their F1 scores are higher than 40%. This shows that the proposed MAP models are well suitable to this few-shot scenario.

4.11 Error Analysis

In this section, we present a few misclassification cases to explore the potential limitations of the proposed MAP model. Note that the case study is perform based on the RoBERTa based MAP model. After careful selection, we highlight four cases here, as each represents a type or scenario where the MAP struggles.

The first occurs when the speaker conceals his facial expression and utters neutral words. In this situation, the MAP model makes the wrong decision as it cannot incorporate the visual information. Thus, the decision relies solely on the textual words. The second example demonstrates that the MAP has difficulty distinguishing between similar emotions, such as disgust and surprise. Both emotions are expressed through exaggerated expressions and a changing tone, leading to incorrect predictions.

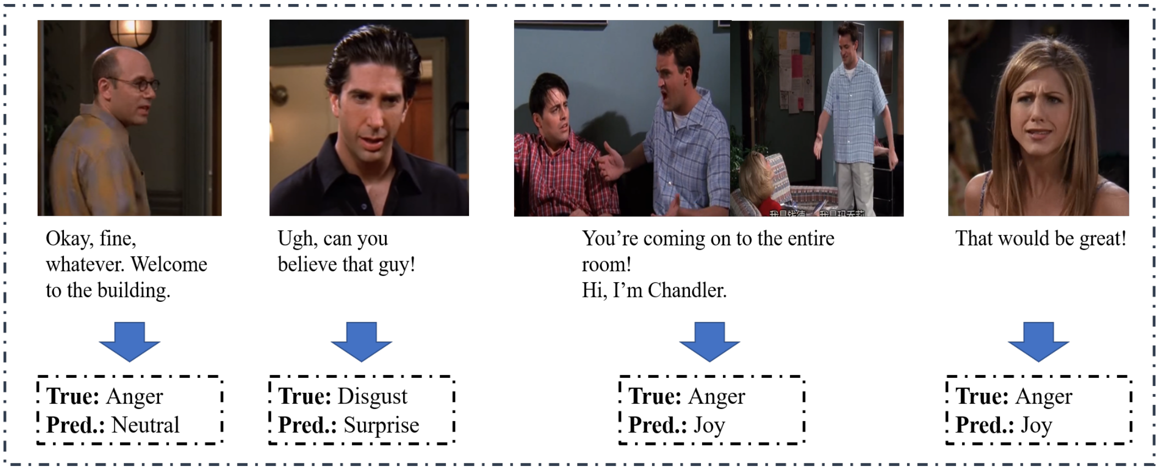

Figure 5: A few misclassification examples.

The third case presents a peculiar phenomenon with two different scenarios in one utterance. The same speaker in two scenarios expresses different emotions, such as anger toward to a friend and then joy upon seeing a cute girl, all within one utterance. The MAP predicts one emotion correctly but misses the other one. The fourth case represents sarcastic scenarios, where the speaker conveys a sarcastic attitude by combing positive words and tones with negative facial expressions. As a result, the MAP model makes the wrong decision, suggesting that further refinement is needed to deal with the sarcastic phenomenon.

5. Ethical Considerations

In this section, we briefly present the ethical considerations associated with each step of the development process, e.g., task, method and data (Mohammad, 2022).

Our research on multi-modal prompt learning for few-shot emotion recognition in conversations (ERC) is guided by ethical considerations throughout the study. We recognize the importance of ethical conduct in data collection, model development, and potential societal impact. The following points highlight the ethics considerations associated with our work:

Task. Emotion recognition in conversations involves analyzing and interpreting individuals' emotions, which can touch upon personal experiences, traumas, or sensitive topics. Researchers must consider the potential emotional impact on participants during data collection, annotation, and analysis. It is essential to prioritize the emotional well-being and mental health of the individuals involved and to handle the data and findings with sensitivity and care.

Emotion recognition in conversations often involves the use of automated systems and machine learning algorithms. It is important to acknowledge the limitations and potential biases of these systems and consider the ethical implications of their decisions. Transparency, explainability, and ongoing evaluation of the system's performance can help ensure that the automated analysis is fair, reliable, and accountable.

Data. (1) Data Collection: The data used in our study are sourced from publicly available benchmark datasets, namely MELD and IEMOCAP. These datasets have been previously released and annotated with consent from the participants. We acknowledge the efforts of the original data collectors in obtaining necessary permissions and ensuring compliance with ethical guidelines.

(2) Privacy and Anonymity: The data we used for model training and evaluation are anonymized and do not contain personally identifiable information. We are committed to preserving the privacy and confidentiality of individuals whose conversations are represented in the dataset.

(3) Informed Consent: The original data collection processes for MELD and IEMOCAP involved obtaining informed consent from the participants. However, as researchers using pre-existing datasets, we did not directly interact with the participants nor had the opportunity to obtain additional consent. We respect the consent provided by the original data collectors and comply with their ethical protocols.

Method. (4) Bias Mitigation: We acknowledge the potential presence of biases in the labeled data used for training our model. While we have made efforts to minimize biases during the data collection and annotation processes, it is important to note that biases may still exist, reflecting societal and cultural influences present in the data. We encourage future researchers to address bias mitigation techniques and ensure fairness in the development and deployment of emotion recognition models.

(5) Responsible Use: As researchers, we advocate for the responsible use of emotion recognition technology. While our proposed method aims to enhance few-shot emotion recognition performance, we emphasize the need for ethical considerations in real-world applications. It is essential to deploy such models in ways that respect user privacy, avoid harmful consequences, and consider potential biases and limitations.

(6) Transparency and Explainability: As our MAP framework leverages deep learning techniques, ensuring transparency and explainability is important. Model transparency enables stakeholders to understand how the system works, how prompts are generated, and how the model arrives at its predictions. Providing explanations for model outputs can foster trust and enable users to understand and question the system's decisions.

By incorporating these ethics considerations into our research, we aim to ensure transparency, accountability, and responsible practices in the field of emotion recognition in conversations.

6. Conclusions and Future Work

Emotion Recognition in Conversation (ERC) presents a fascinating and complex challenge in the fields of Natural Language Processing (NLP) and Artificial Intelligence (AI). Most current research has predominantly focused on fine-tuning pre-trained language models (PLMs) using a plethora of labeled samples to improve classification performance. In contrast, we propose a Multi-modal Attention Prompt (MAP) learning framework for few-shot ERC. By creating multiple multi-modal prompts that integrate both multi-modal and contextual information, our model effectively addresses the few-shot ERC problem. Our comprehensive experiments demonstrate the efficacy of our proposed model, outperforming current state-of-the-art ERC models and showcasing significant improvements in micro F1 scores. Moreover, our experiments highlight the potential and flexibility of multi-modal prompt learning in ERC tasks. In this work, we focus on the manual design of prompts instead of constructing learnable soft prompts, which lays the foundation for future exploration in this direction.

As we progress, our plans encompass the development of more efficient and sophisticated prompts, as well as an extension of the application of the MAP model to other multi-modal tasks. These tasks may include sentiment analysis, sarcasm detection, and multi-modal question answering. Moreover, we aspire to delve into automated methods for prompt generation and fusion, which hold the potential to further enhance our model's performance and make significant contributions to the broader field of Emotion Recognition and Natural Language Understanding.

It's important to note that both manual and automatic prompt generation approaches possess their distinct advantages and limitations. Manual prompts, crafted by human designers, bring expertise and domain knowledge to ensure the quality and relevance of the prompts. Nevertheless, this manual process can be time-consuming and may not scale effectively to accommodate large datasets or diverse domains. On the other hand, automatic prompt generation offers the advantages of adaptability and scalability, enabling the model to generate prompts tailored to specific contexts or tasks. However, the effectiveness of automatic prompt generation methods hinges upon the quality of the training data and the learning algorithms utilized.

In our forthcoming work, we have set the objective of exploring and developing both manual and automatic prompt generation approaches, aiming to enhance prompt efficiency and sophistication. This involves refining and optimizing the design of manual prompts based on empirical insights and user feedback. Additionally, we will investigate techniques for automated prompt generation to adaptively create prompts that improve model performance. Our ultimate goal is the ongoing enhancement of prompt quality and effectiveness for emotion recognition in conversational contexts.

Furthermore, it's worth noting that the proposed MAP method has the potential for application to more potent large language models, such as LLaMA/LLaMA 2. However, this direction will be reserved for our future work.

Limitations. There are also a few limitations. (1) The experiments conducted in this study were primarily focused on two benchmark datasets, MELD and IEMOCAP. It is important to consider that the performance and generalizability of the MAP framework to other datasets. (2) The MAP framework utilizes six manually-designed multi-modal prompts. While these prompts were constructed based on intuition and considerations of different modalities, there is a possibility that other prompt variations or combinations could potentially yield different results. The exploration of alternative prompt designs or automated prompt generation methods could further enhance the framework's performance. It is essential to acknowledge these limitations to provide a clear context for the proposed model and to inspire our further investigation and improvements in the field of emotion recognition in conversations.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China 62176076, Natural Science Foundation of Guang Dong 2023A1515012922, Shenzhen Foundational Research Funding JCYJ20220818102415032, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 022B1212010005, and the Major Key Project of PCL PCL2023A09.

Appendix. The Synonyms List

The detailed synonyms list is shown in Table 8. We will map the synonyms into the corresponding labels.

Emotions	**Synonyms**
Anger	anger, angry, annoyed, furious, bitter, acrimony, animosity, annoyance, antagonism, displeasure, enmity, exasperation, fury, hatred, impatience, indignation, ire, irritation, outrage, passion, rage, resentment, temper, violence, chagrin, choler, conniption, dander, disapprobation, distemper, gall, huff, infuriation, irascibility, irritability, miff, peevishness, petulance, pique, rankling, soreness, stew, storm, tantrum, tiff, umbrage, vexation

Disgust	disgust, distaste, antipathy, aversion, dislike, distaste, hatred, loathing, repulsion, revulsion, abhorrence, abomination, detestation, hatefulness, nausea, objection, repugnanc, revolt, satiation, satiety, sickness, surfeit, nauseation, nauseousness
Fear	fear, alarm, angst, anxiety, apprehension, awe, concern, despair, dismay, doubt, dread, horror, jitters, panic, scare, suspicion, terror, unease, uneasiness, worry, abhorrence, agitation, apprehensivenes, aversion, consternation, cowardice, creeps, discomposure, disquietude, distress, faintheartedness, fearfulness, foreboding, fright, funk, misgiving, nightmare, phobia, presentiment, qualm, reverence, revulsion, timidity, trembling, trepidation, recreancy
Joy	joy, happy, joyful, amusement, bliss, charm, cheer, comfort, delight, elation, glee, humor, pride, satisfaction, wonder, alleviation, animation, delectation, diversion, ecstasy, exultation, exulting, felicity, festivity, frolic, fruition, gaiety, gem, gladness, gratification, hilarity, indulgence, jewel, jubilance, liveliness, luxury, merriment, mirth, prize, rapture, ravishment, refreshment, rejoicing, revelry, solace, sport, transport, treasure, treat, good humor, pride, joy, regalement
Neutral	neutral, neuter, litmusless, disinterested, evenhanded, inactive, indifferent, nonaligned, nonpartisan, unbiased, uncommitted, undecided, unaligned, unconcerned, unprejudiced
Sadness	sad, sadness, anguish, grief, heartache, heartbreak, hopelessness, melancholy, misery, mourning, poignancy, sorrow, blahs, bleakness, bummer, cheerlessness, dejection, despondency, disconsolateness, dispiritedness, distress, dolefulness, dolor, downer, dysphoria, forlornness, funk, gloominess, letdown, listlessness, moodiness, mopes, mournfulness, sorrowfulness, tribulation, woe, blue devils, blue funk, broken heart, dismals, downcastness, grieving, heavy heart, the blues, the dumps
Surprise	surprise, surprised, surprising, amazement, astonishment, awe, bewilderment, consternation, curiosity, disappointment, jolt, miracle, revelation, shock, wonder, abruptness, attack, bombshell, disillusion, epiphany, eureka, fortune, godsend, incredulity, kick, marvel, miscalculation, phenomenon, portent, precipitance, precipitation, unexpected, unforeseen

Table 8: A list of all synonyms

References

Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 258–266, Marseille, France. European Language Resources Association.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 325–335.

Chuang, Z.-J., & Wu, C.-H. (2004). Multi-modal emotion recognition from speech and text. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, pp. 45–62.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Deng, S., Shi, G., Wu, L., Xing, L., Hu, W., Zhang, H., & Xiang, Y. (2022a). Simemotion: A simple knowledgeable prompt tuning method for image emotion classification. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part III*, pp. 222–229. Springer.

Deng, S., Wu, L., Shi, G., Xing, L., & Jian, M. (2022b). Learning to compose diversified prompts for image emotion classification. *arXiv preprint arXiv:2201.10963*, *n/a*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186.

Fan, Y., Lu, X., Li, D., & Liu, Y. (2016). Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pp. 445–450.

Fei, Z., Yang, E., Li, D. D.-U., Butler, S., Ijomah, W., Li, X., & Zhou, H. (2020). Deep convolution network based emotion analysis towards mental health care. *Neurocomputing*, *388*, 212–227.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. F. (2019). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 154–164.

Gluz, J., & Jaques, P. A. (2017). A probabilistic formalization of the appraisal for the occ event-based emotions. *Journal of Artificial Intelligence Research*, *58*, 627–664.

Graves, A., Jaitly, N., & Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pp. 273–278. IEEE.

Gu, Y., Han, X., Liu, Z., & Huang, M. (2022). PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Han, W., Jiang, T., Li, Y., Schuller, B., & Ruan, H. (2020). Ordinal learning for emotion recognition in customer service calls. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6494–6498. IEEE.

Hu, J., Liu, Y., Zhao, J., & Jin, Q. (2021). Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 5666–5675, Dublin, Ireland. Association for Computational Linguistics.

Huang, Y., Qian, K., & Yu, Z. (2022). Learning a better initialization for soft prompts via meta-learning. In *The 29th International Conference on Computational Linguistics*, pp. 287–288, Gyeongju. Association for Computational Linguistics.

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2019). A survey of computational approaches and challenges in multimodal sentiment analysis. *Int J Comput Sci Eng*, *7*(1), 876–883.

Ishiwatari, T., Yasuda, Y., Miyazaki, T., & Goto, J. (2020). Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7360–7370.

Jia, A., He, Y., Zhang, Y., Uprety, S., Song, D., & Lioma, C. (2022). Beyond emotion: A multi-modal dataset for human desire understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1512–1522.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kollias, D., & Zafeiriou, S. (2020). Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, *12*(3), 595–606.

Le, H. T., & Vea, L. A. (2016). A customer emotion recognition through facial expression using kinect sensors v1 and v2: A comparative analysis. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, pp. 1–7.

Li, C., Bao, Z., Li, L., & Zhao, Z. (2020). Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition. *Information Processing & Management*, *57*(3), 102185.

Li, X., Li, J., Zhang, Y., & Tiwari, P. (2021). Emotion recognition from multi-channel eeg data through a dual-pipeline graph attention network. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3642–3647. IEEE.

Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., Zhao, Z., Kumar, N., & Marttinen, P. (2022a). Eeg based emotion recognition: A tutorial and review. *ACM Computing Surveys*, *55*(4), 1–57.

Li, Z., Tang, F., Zhao, M., & Zhu, Y. (2022b). EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1610–1618, Dublin, Ireland. Association for Computational Linguistics.

Liang, X., Zou, Y., Zhuang, X., Yang, J., Niu, T., & Xu, R. (2023). Mmateric: Multi-task learning and multi-fusion for audiotext emotion recognition in conversation. *Electronics*, *12*(7), 1534.

Liu, X., You, J., Wu, Y., Li, T., Li, L., Zhang, Z., & Ge, J. (2020). Attention-based bidirectional gru networks for efficient https traffic classification. *Information Sciences*, *541*, 297–315.

Liu, Y., Li, Q., Wang, B., Zhang, Y., & Song, D. (2023). A survey of quantum-cognitively inspired sentiment analysis models. *ACM Computing Surveys*, *56*, 1–37.

Liu, Y., Zhang, Y., Li, Q., Wang, B., & Song, D. (2021). What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 871–880.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 1–10.

Lu, X., Zhao, Y., Wu, Y., Tian, Y., Chen, H., & Qin, B. (2020). An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4078–4088. IEEE.

Ma, H., Wang, J., Lin, H., Pan, X., Zhang, Y., & Yang, Z. (2022). A multi-view network for real-time emotion recognition in conversations. *Knowledge-Based Systems*, *236*, 107751.

Ma, J., Rong, L., Zhang, Y., & Tiwari, P. (2023). Moving from narrative to interactive multimodal sentiment analysis: A survey. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *24*(1), 1–20.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 6818–6825.

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1–17.

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pp. 26–34.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pp. 201–237. Elsevier.

Mohammad, S. M. (2022). Ethics sheets for ai tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8368–8379, Dublin, Ireland. Association for Computational Linguistics.

Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301–326.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436–465.

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, *51*(4), 480–499.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 527–536.

Ribeiro, B., Oliveira, G., Laranjeira, A., & Arrais, J. P. (2017). Deep learning in digital marketing: brand detection and emotion recognition. *International Journal of Machine Intelligence and Sensory Signal Processing*, *2*(1), 32–50.

Rozgić, V., Ananthakrishnan, S., Saleem, S., Kumar, R., & Prasad, R. (2012). Ensemble of svm trees for multimodal emotion recognition. In *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference*, pp. 1–4. IEEE.

Saxena, A., Khanna, A., & Gupta, D. (2020). Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, *2*(1), 53–79.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4222–4235, Online. Association for Computational Linguistics.

Song, L., Zhang, Y., & Hou, Y. (2018a). Convolutional neural network with pair-wise pure dependence for sentence classification. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 117–121. IEEE.

Song, L., Liu, J., Qian, B., Sun, M., Yang, K., Sun, M., & Abbas, S. (2018b). A deep multimodal cnn for multi-instance multi-label image classification. *IEEE Transactions on Image Processing*, *27*(12), 6025–6038.

Spezialetti, M., Placidi, G., & Rossi, S. (2020). Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 145.

Tu, G., Wen, J., Liu, C., Jiang, D., & Cambria, E. (2022). Context-and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence*, *3*, 699–708.

Wang, J., Ma, F., Zhang, Y., & Song, D. (2022a). A multibias-mitigated and sentiment knowledge enriched transformer for debiasing in multimodal conversational emotion recognition. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 499–512. Springer.

Wang, Z., Wang, P., Liu, T., Cao, Y., Sui, Z., & Wang, H. (2022b). Hpt: Hierarchy-aware prompt tuning for hierarchical text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3740–3751, Abu Dhabi. Association for Computational Linguistics.

Wu, H., & Shi, X. (2022a). Adversarial soft prompt tuning for cross-domain sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2438–2447.

Wu, H., & Shi, X. (2022b). Adversarial soft prompt tuning for cross-domain sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2438–2447.

Xiao, Y., Zhao, H., & Li, T. (2020). Learning class-aligned and generalized domain-invariant representations for speech emotion recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *4*(4), 480–489.

Xu, S., Pang, L., Shen, H., & Cheng, X. (2022). Match-prompt: Improving multi-task generalization ability for neural text matching via prompt learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 2290–2300.

Yang, X., Feng, S., Wang, D., Hong, P., & Poria, S. (2023). Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11575–11589, Toronto, Canada. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 5753–5763, Vancouver, Canada. NIPS.

Yi, J., Yang, D., Yuan, S., Cao, C., Zhang, Z., & Xiao, Y. (2022). Contextual information and commonsense based prompt for emotion recognition in conversation. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 1338–1345, Grenoble, France. ECML.

Zhang, Y., Jia, A., Wang, B., Zhang, P., Zhao, D., Li, P., Hou, Y., Jin, X., Song, D., & Qin, J. (2023). M3gat: A multi-modal multi-task interactive graph attention network for conversational sentiment analysis and emotion recognition. *ACM Transactions on Information Systems*, *42*(1), 1–32.

Zhang, Y., Li, Q., Song, D., Zhang, P., & Wang, P. (2019). Quantum-inspired interactive networks for conversational sentiment analysis. In *Proceedings of the Twenty-Eighth*

International Joint Conference on Artificial Intelligence IJCAI-19, pp. 5436–5442. International Joint Conferences on Artificial Intelligence Organization.

Zhang, Y., Li, X., Rong, L., & Tiwari, P. (2021a). Multi-task learning for jointly detecting depression and emotion. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3142–3149. IEEE.

Zhang, Y., Liu, Y., Li, Q., Tiwari, P., Wang, B., Li, Y., Pandey, H. M., Zhang, P., & Song, D. (2021b). Cfn: A complex-valued fuzzy network for sarcasm detection in conversations. *IEEE Transactions on Fuzzy Systems*, *29*(1), 3696–3710.

Zhang, Y., Rong, L., Li, X., Tiwari, P., Zheng, Q., & Liang, H. (2021c). Medseq2seq: A medical knowledge enriched sequence to sequence learning model for covid-19 diagnosis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3181–3184. IEEE.

Zhang, Y., Song, D., Li, X., & Zhang, P. (2018). Unsupervised sentiment analysis of twitter posts using density matrix representation. In *European Conference on Information Retrieval*, pp. 316–329.

Zhang, Y., Song, D., Li, X., Zhang, P., Wang, P., Rong, L., Yu, G., & Wang, B. (2020). A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Information Fusion*, *62*, 14–31.

Zhang, Y., Song, D., Zhang, P., Li, X., & Wang, P. (2019). A quantum-inspired sentiment representation model for twitter sentiment analysis. *Applied Intelligence*, *49*(8), 3093–3108.

Zhang, Y., Tiwari, P., Rong, L., Chen, R., AlNajem, N. A., & Hossain, M. S. (2021a). Affective interaction: Attentive representation learning for multi-modal sentiment classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *136*, 1–23.

Zhang, Y., Tiwari, P., Song, D., Mao, X., Wang, P., Li, X., & Pandey, H. M. (2021b). Learning interaction dynamics with an interactive lstm for conversational sentiment analysis.. *Neural Networks*, *133*, 40–56.

Zhang, Y., Tiwari, P., Zheng, Q., El Saddik, A., & Hossain, M. S. (2022). A multimodal coupled graph attention network for joint traffic event detection and sentiment classification. *IEEE Transactions on Intelligent Transportation Systems*, *24*, 8542–8554.

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S., & Sun, X. (2014). Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 47–56.

Zhou, Y., Kang, X., & Ren, F. (2023). Prompt consistency for multi-label textual emotion detection. *IEEE Transactions on Affective Computing*, 1–13.