Master's Thesis

A Text Prompt-based Fine-tuning Method for Multimodal Sentiment Analysis

TANG WENNA

Supervisor SHOGO OKADA

Graduate School of Information Science
Japan Advanced Institute of Science and Technology
(Master Degree)

March, 2023

Abstract

The accelerating evolution of societal dynamics has brought forth an increasingly diverse array of information types. Of particular significance is the burgeoning interest in sentiment analysis, spurred by its versatile applications across various domains. The discernment of sentiments holds particular relevance, given its potential utility in numerous applications. Consequently, there has been a concerted effort to delve into the intricacies of multiple modalities to unearth latent information. This has given rise to a spectrum of methodologies aimed at effectively handling the complexities inherent in the amalgamation of diverse modalities. This has given rise to a spectrum of methodologies aimed at effectively handling the complexities inherent in the amalgamation of diverse modalities.

Concurrently, the societal discourse on mental health has manifested in an upsurge of applications pertaining to sentiment analysis and emotion detection. This evolving landscape has witnessed the trajectory of sentiment analysis tasks, progressing from unimodal and bimodal to the contemporary trimodal paradigm. The concomitant escalation in the demand for adeptly managing multiple modalities has been a discernible trend in recent years. Within this overarching milieu, this research introduces a text prompt-based fine-tuning method designed to address the challenges posed by distinct modalities within the framework of multimodal sentiment analysis.

The research objective is the pursuit of an interpretable and simplified approach for alleviate the gap between disparate modalities in a natural language manner. In this pursuit, an initial recourse is made to a prompt-based methodology during the fine-tuning phase. This methodological choice is grounded in its transformative capacity, recasting downstream tasks as cloze-filling exercises—a format inherently conducive to enhanced human comprehension. However, the matter lies in generating semantically rich representation from modalities beyond textual data.

To achieve this goal, a text prompt-based fine-tuning method is proposed in this research. This approach hinges on the meticulous application of manually crafted rules to generate textual descriptions from visual and auditory modalities. Consequently, the semantic descriptions is combined with textual information in a natural language fomrat with a fixed template. Due to its interpretability in natural language, this method is capable to understand by human beings. In other words, it also is able to make an adaption to different task. Subsequently, the process entails the formulation of a prompt function,

which is fed into a pre-trained language model and make the prediction. In the validation of this methodology, experiments are conducted leveraging the MELD dataset. Comparative analyses juxtaposing baseline results with an augmented baseline featuring attention mechanisms underscore the efficacy of the proposed method.

In conclusion, this research propose a method applying with the prompt-based fine-tuning method to navigate the intricate landscape of multimodal sentiment analysis. The fusion method between different modalities of interpretability and simplification is shown.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Within the ever-expanding realm of big data, which encompasses a myriad of information types such as audio, image, and text. Besides, the potential to uncover hidden insights grows exponentially. This expansion has concurrently led to the emergence and development of the research field in Multimodal Sentiment Analysis (MSA). This specialized field addresses the complexities introduced by the coexistence of various modalities and finds application in diverse fields such as opinion mining[1, 2], depression detection[3, 4], recommendation systems[5], and more.

The task of sentiment analysis initially emerged in the research field of natural language processing, primarily relying on textual modalities for analysis. Over time, this task evolved from an unimodal modality to encompassing multiple modalities. This shift was based on the assumption that exploring various cues could contribute significantly to sentiment analysis within a multimodal scenario. In MSA tasks, addressing the gap between multiple modalities emerges as a primary challenge. The most prevalent modalities include text (e.g., spoken words), audio (e.g., pitch, tone, intensity), and image (e.g., facial expression, eye gaze). Each modality is accompanied by its unique data format. Consequently, when dealing with multiple modalities simultaneously, direct combination of raw data and feeding it into a model is not feasible. Fusion work becomes imperative, aiming to project these diverse modalities into a unified feature space – a common vector space that the model can recognize.

To tackle the fusion issue, extensive efforts towards developing fusion networks within model architectures are required to effectively combine different modalities. Conventional approaches to address this challenge include a number of fusion techniques, such as early fusion[6], late fusion[7], and tensor fusion[8]. Early fusion initially combines all features from each modality into a single feature vector, followed by the application of a classification al-

gorithm. This method ensures the early identification of correlations between multiple modalities but comes with the disadvantage of lacking the representation of intra-modality dynamics. On the other hand, late fusion involves fusion at a later stage, specifically during the classification stage. Prior to this stage, a dedicated model is designed for each modality. However, this approach has the drawback of being time-consuming as different models must be trained separately for each modality. Tensor fusion represents an approach that employs the model to learn interactions from unimodal, bimodal, and trimodal data. While each of these approaches has its own merits and drawbacks, a common issue persists—learning the correlation between different modalities.

Although these methodologies have demonstrated enhanced performance in downstream classification tasks, they still exhibit common limitations. Conventional fusion techniques predominantly require the implementation of a neutral fusion layer to learn the relationship between different modalities with updating numbers of parameters for model. Hence, with the aim of alleviating the burden on the fine-tuning phase on downstream tasks while effectively harnessing the capabilities of pre-trained models, we advocate for a more simplified and interpretable approach in MSA tasks. Furthermore, our current focus is on uncovering the interpretability of models, given that models often appear as black boxes, making them unclear for human comprehension, especially with the gradual complexity of deep neural networks[9]. There is a growing need to discern the rationale behind the decisions made by models. Within real conversations, humans typically gauge the sentiments of others through voice tones, facial expressions, and spoken words. These human judgments are intuitive, contrasting with the minuscule features detected by neural networks. The surge in research emphasizes the importance of interpretability in multimodal analysis[10]. For instance, [10] introduced a method that analyzes textual explanations as counterfactual explanations derived from images. It is crucial to note that interpretability predominantly centers on textual information, as text inherently provides clear and explicit explanatory features from the perspective of natural language. To facilitate the model's understanding of this process, we posit that the conversion between modalities other than text and text can offer humans an intuitive and lucid comprehension of the model functioning.

Aligns with the rapid development of large language models (LLMs), such as GPT-3[11] and LLaMA[12], Prompt-based method is proposed as a essential product as a recent advanced method in the research field of natural language processing(NLP). Prompt-based method involves providing prompts to Language Models (LMs) and having them make predictions for specific words in a Masked Language Model (MLM). This approach transforms the

downstream task into a cloze-filling task, offering improved adaptability to new scenarios compared to traditional fine-tuning methods[13]. The use of prompts makes the process more explainable for human comprehension, providing clearer guidance in a natural language manner. In response to the merits of prompts, the application of this method in MSA tasks has been vigorously promoted in recent years[14, 15]. Indeed, within the MSA research domain, enhancing the applicability of models to wild data poses a significant challenge. The utilization of prompts, in contrast to traditional methods, offers the advantage of achieving improved applicability without the necessity of updating a substantial number of model parameters. Instead, this can be accomplished by altering the prompt embedding with less parameter update in fine-tuning step.

Leveraging the advantages of the prompt-based method, this study aims to apply it to MSA tasks in an interpretable and simplified manner. In contemplating this, natural language descriptions are deemed essential for interpretability. Considering the text-dominant phenomenon[16], which suggests that the most substantial contribution to the final result originates from the text modality, we posit that integrating information from other modalities through the translation to textual descriptions is a promising avenue in MSA. In essence, this involves adding visual and acoustic cues by generating semantic descriptions and fusing them into a multimodal prompt for fine-tuning PLM. In light of these considerations, we introduce a text prompt-based tuning approach through the way of generating textual information from extracted features in a manual setting rules in this study.

Within the scope of this research, we propose a text prompt-based fine-tuning method with semantic descriptions for MSA tasks. The objective of this study is to elucidate a simplified and interpretable method for modality fusion. Through the generation of textual descriptions from both acoustic and visual modalities, our method aims to bridge the gap between different modalities, thereby optimizing the utilization of LLM.

Chapter 2

Related Works

2.1 Multimodal Sentiment Analysis

For multimodal sentiment classification tasks, the focal point often revolves around the fusion technique, a topic frequently discussed within the literature. The overarching objective of MSA tasks is to extract latent cues from diverse modalities and fuse them cohesively to predict sentimental labels, such as positive, negative, and neutral. The efficacy of the extraction method is indispensable when dealing with disparate data types. To address the fusion challenge, a multitude of fusion techniques have been proposed in recent years. This evolutionary progression extends from uni-modal (text) methodologies to bi-modal (combinations of text, audio, and visual) approaches, culminating in tri-modal (text-audio-visual) models.

Concurrently, the majority of fusion studies have predominantly focused on the application of enhanced attention-based or Long Short-Term Memory (LSTM)-based components. These components are considered state-of-the-art mechanisms for handling multiple modalities, given their proficiency in unraveling intricate interactions. Therefore, through the introduction of these models, we can gain a brief understanding of how conventional networks operate. In the case of attention-based methods, the work of MARN[6] stands out, wherein researchers proposed multi-attention-based neural components to align features from diverse modalities. In [17], the authors employed an ABS-LSTM structure to generate local and global embeddings. Additionally, [18] proposed a Bi-LSTM, coupled with an attention model, adeptly extracts contextual information from utterances, leveraging both types of components. Undoubtedly, the conventional fusion networks mentioned above consistently yield commendable performance across various datasets[19, 20, 21]. However, it is imperative to acknowledge that these enhanced components

necessitate significant computing resources for learning how to align multimodal features. Furthermore, when it comes to feature alignment, discerning which specific components contribute most significantly to the final classification result poses a substantial challenge. The intricacies of detecting the pivotal contributors to the final classification outcome underscore the need for further research and methodological refinements in the domain of feature alignment within MSA tasks.

2.2 Prompt-based Method

Prompting, originally a popular topic in natural language processing (NLP), has gained renewed attention due to advancements in large language models. The prompt-based method capitalizes on the capabilities of pre-trained language models, enabling the downstream task to model the probability of text directly[13]. In comparison with conventional fine-tuning methods, it offers greater parameter efficiency and eliminates the need to train additional layers for the downstream task. The prompt-based method transforms the classification task into a cloze-filling problem, introducing prompts with a slot to be completed by a masked language model(MLM). From [13], two pivotal components characterize the prompt-based method: template and verbalizer. The impact of different templates and verbalizers on results has been emphasized by [22].

Extensive efforts have been devoted to automatically generating templates and verbalizers, as explored by LM-BFF[23], AutoPrompt[24], and LAMA[25]. Studies from [26] have delved into the comparison of manually picked prompts with automated ones, concluding that manually selected prompts may not achieve the performance of automated alternatives. For example, [27] aided language models in understanding a given task through the approach that involves rephrasing input examples into cloze-style phrases. This consideration underlies our approach in this research, where we employ both manual and automated prompts to strike a balance in results.

The application of the prompting-based method extends into the research domain of MSA. Several studies, such as [28], have showcased its effectiveness in MSA tasks. In their work, [28] employed multimodal prompts, utilizing the NF-ResNet architecture [29] to project image representations into text feature space and leveraging ClipCap [30] to generate semantic descriptions for images. Similarly, [31] utilized ResNet [32] to generate image embeddings, subsequently fusing them with textual modality. These approaches commonly leverage neural networks to project image representations into text feature space. However, the persisting issue of interpretability remains

a challenge in these models. In particular, existing models, including those referenced above, often include exact embeddings with tokens, resulting in a lack of interpretability. The inclusion of non-textual information alongside textual data poses challenges in adopting the prompt-based method in a fully interpretable natural language manner.

To address this interpretability issue, our research explores the possibility of a simplified and interpretable approach to fusing multiple modalities based on natural language templates. This approach aims to overcome the challenges posed by current models and enhance the interpretability of multimodal sentiment analysis systems.

Chapter 3

Methodology

3.1 Task Formulation

Applying with prompt-based method in MSA, the classification task is generally formatted as follows. Access to a PLM, denoted as M, with a constructed prompt function $f_{prompt}(\cdot)$, template T with a masked slot and a verbalizer(label mapping) space V. Both the template and verbalizer could be generated by MLM automatically or manually. The output is obtained from:

$$M(f_{prompt}(x_i, T, L)) \tag{3.1}$$

where x_i indicates the i^{th} sample from train dataset D_{train}.

3.2 Proposed Method

In this research, our goal is to fine-tune the PLM M for the multimodal sentiment classification task on the prompt-based approach, the task is formulated in the last subsection. The whole architecture is illustrated in Figure 3.1. There are two primary phases:1) Constructing appropriate prompt function with auto verbalizer searching; 2) Making the prediction with the prompt-based fine-tuning method.

3.2.1 Data Processing

To prepare for the subsequent stages of constructing prompts, the initial step involves processing all raw data available in textual format. Specifically, for textual information, the primary selection consists of opting for the original utterance sentences directly from the dataset. This approach ensures that the foundation of the dataset is rooted in the authentic expressions and linguistic

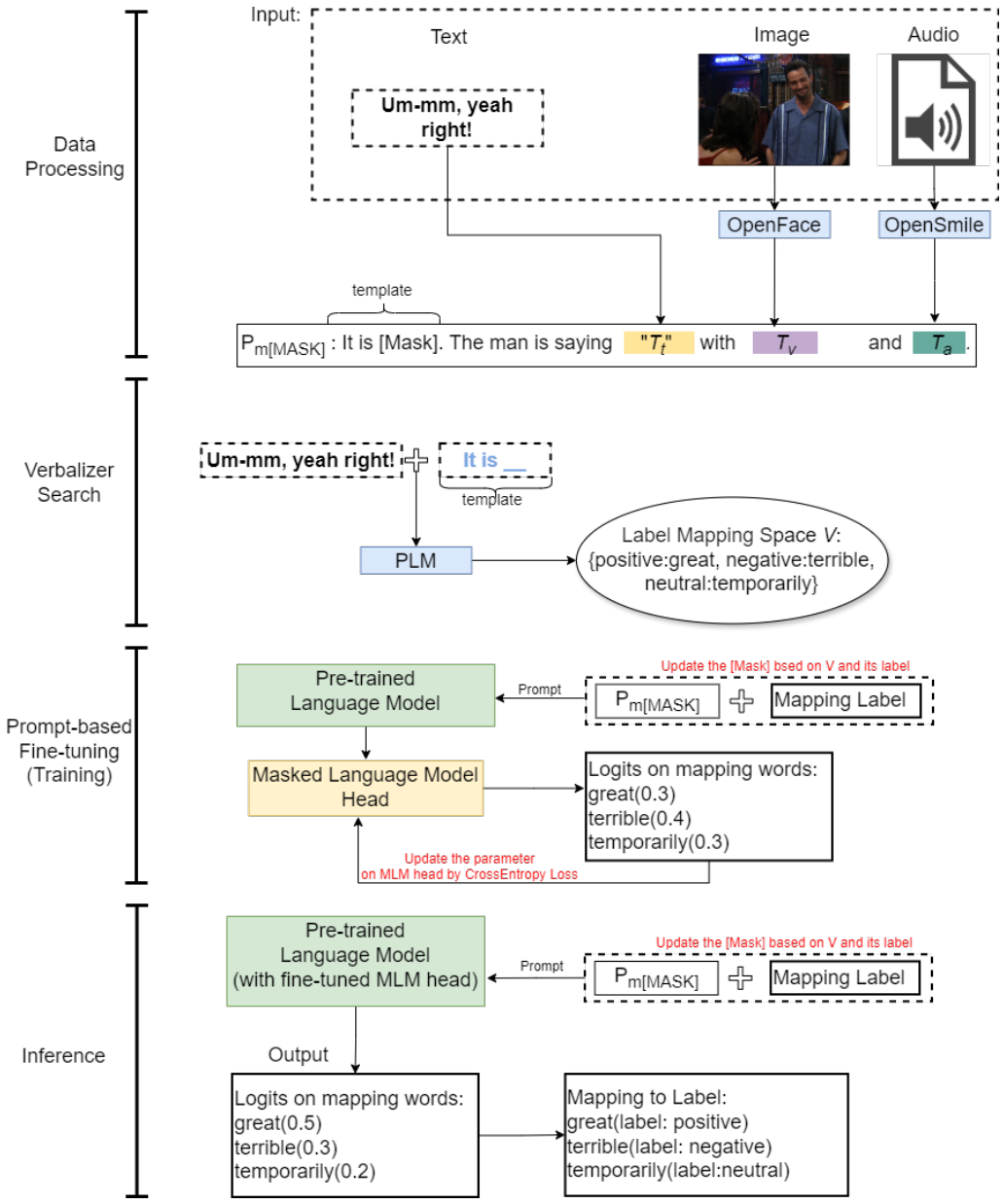

Figure 3.1: : Three primary steps are involved in: 1) Data processing. Manually setting rules for generating textual information from audio features and visual features. 2) Verbalizer Search: Searching appropriate label mapping space for multimodal prompt. Specifically, the label mapping positive:great, negative: terrible, neutral: temporarily is solely an exmaple for clarify the process, not the true label space utilized in this research. 3) Prompt-based Fine-tuning: During the fine-tuning step, we first feed the constructed prompt, with replacements from the [Mask] slot based on the label mapping space generated in the verbalizer search step, into PLM to update the MLM head applying with the CrossEntropy loss function. 4) Inference: multimodal prompt with [Mask] slot is fed into PLM with updated MLM head, then predict the logits for mapping words and project them back to label.

nuances present in the raw data. Moving beyond text, for other modalities, a crucial undertaking involves the manual generation of semantic descriptions. This manual rule-based generation process aims to extract the essence and meaning embedded in each modality distinct from the text. By employing a manual rule-based approach, we ensure precision and intentionality in crafting semantic descriptions that align with the inherent characteristics of each modality.

Ensuring comprehensive information from various modalities is included in the prompt is a key step. For visual modality, facial expressions have been underscored their pivotal roles in numerous studies in the research field of MSA. Furthermore, human sentiment is able to be detected by additional cues, such as head pose, facial action unit(AU), and eye gaze. Of particular note, AU presents merit due to its fixed combination and well-defined framework. In this context, we utilize OpenFace[33] to detect valuable visual features for constructing detailed and informative textual descriptions, especially the AUs and its corresponding intensity. In the context of textual description generation from visual modality, the mapping rule comprises two components: AU name and its associated intensity word. Employing an average approach for each sample, the action units are mapped in accordance with the Facial Action Coding System(FACS)[34]. The comprehensive list of AUs and their corresponding mapping names are delineated in Table 3.1.

While the second facet of intensity word translation adheres to a predefined rule as elucidated in Table 3.2. The intensity values detected by OpenFace span a range from 0 to 5, where a higher value signify a more pronounced intensity. Corresponding emotional words are used to express the intensity of each detected AU. Notably, for each value range, the intensity word is randomly selected from the available candidates in the corresponding word group. This method ensures a diversified and contextually appropriate expression of intensity across the detected values.

In examining the acoustic modality, textual descriptions rely on feature extraction through OpenSmile[35]. Initially, we extracted features following the IS09 feature set rule. This choice was motivated by the recognition of the significance of vocal parameters, including pitch and intensity, in emotion detection and sentiment analysis, as emphasized in [36]. These vocal parameters serve as crucial components influencing the accurate interpretation of emotional nuances within spoken content. From the feature extracted on the feature set of IS09, three fundamental elements are selected: pitch, intensity, and zero-crossing rate. The specific values of these features are chosen in an average manner for each video.

AUs	Full Name
AU1	INNER BROW RAISER
AU2	OUTER BROW RAISER
AU4	BROW LOWERER
AU5	UPPER LID RAISER
AU6	CHEEK RAISER
AU7	LID TIGHTENER
AU9	NOSE WRINKLER
AU10	UPPER LIP RAISER
AU12	LIP CORNER PULLER
AU14	DIMPLER
AU15	LIP CORNER DEPRESSOR
AU17	CHIN RAISER
AU20	LIP STRETCHED
AU23	LIP TIGHTENER
AU25	LIPS PART
AU26	JAW DROP
AU28	LIP SUCK
AU45	BLINK

Table 3.1: Action unit and its full name

3.2.2 Prompt Construction

As depicted in 3.1, after generating semantic descriptions T_v for visual modality and T_a for acoustic modality and make a combination between them and textual information from dataset. With a fixed template of *with the intensity of* V_i *and pitch of* V_p *and zero-crossing rate of* V_z , where V_i, V_p, V_z should be replaced by its specific value in an average manner for each video. The manual setting rule governing the combination is formulated as:

$$T_c^i = TP(T_t^i, T_a^i, T_v^i) \tag{3.2}$$

, where i represents the i^{th} sample, T_t denotes the original utterance from dataset, T_v refers to linguistic descriptions for visual features, T_a represents to generated text description for audio features, TP stands for a manual natural language template utilized to seamlessly integrate all the descriptions.

Consequently, to create a comprehensive and fitting prompt for input during the fine-tuning phase, we adopt a prompt-based approach treating multimodal classification as a cloze-filling task. This approach involves two primary steps: template design and answer mapping. Starting with the template, we leverage a general template commonly used in sentiment analysis

Intensity[1]	Mapping Word Candidates
(0,2]	“slightly”,“somewhat”,“a little”,“minially”
(2,3]	”moderately”,”fairly”,”reasonably”,”quite”,”in part”
(3-5]	”extremely”,”intensely”,”passionately”,”overwhelmingly” ”exceedingly”,”profoundly”,”fiercely”

[1] Specifically, '(' denotes open interval, ']' denotes close interval.

Table 3.2: Words mapping rules for AU intensity.

tasks: ”It is.”. This template serves as the foundational structure for constructing prompts, providing a consistent and adaptable framework. In terms of the second step of answer mapping, the objective is to establish a suitable label mapping space from the downstream task's label space to the specific vocabulary of the PLM. In essence, this step entails replacing the [*mask*] position in the prompt with individual vocabulary. Drawing inspiration from recent studies of prompt tuning, such as LM-BFF[23] and AutoPrompt[24]. These studies applied with various of ways to do the label mapping search in PLM automatically. In this study, we employ a label searching approach inspired by the concept from LM-BFF. To mitigate the risk of overfitting and enhance generalization, we exclusively input textual prompts—solely derived from the utterances—using the uniform template '*It is*' into the PLM. This step aims to identify a suitable label space for the prompting process.

The initial phase involves the construction of the label mapping space, denoted as V. This space is fashioned by a pruned set, combined with the top k vocabulary words. The selection criteria for these words are based on the evaluation of their conditional likelihood using the initial label set L. The process of searching for the label space V for each class c belonging to the original label set L is formulated as follows:

$$\mathrm{Top}(k)_{v \in V} \left\{ \sum_{x_i \in D^c_{\mathrm{train}}} \log P_L([\mathrm{Mask}] = v \mid \mathrm{Template}(x_i)) \right\}, \tag{3.3}$$

where k is set to 10 in this research, x_i represents the i^{th} sample, L signifies the initial label space, and *Template* denotes the combination of the original utterance with the template '*It is*' and the masked slot. Subsequently, we fine-tune all assignments, and based on the dev dataset D_{dev}, we rerank to determine the best one. The resultant configuration is utilized as the label mapping space V. This method, inspired by LM-BFF, proves to be highly effective in searching for the label mapping space, offering significant utility in the subsequent steps of constructing the prompt function.

Finally, we construct prompt function as:

$$P_m(T_c^i) = It\ is\ [Mask] + T_c^i \tag{3.4}$$

where $P_m(\cdot)$ denotes the prompt function applied in, *"It is"* is a general template for prompt in sentiment classification task and is fixed in this research, $[Mask]$ should be replaced by specific word in label mapping space generated by the language model or designed manually.

3.2.3 Prompt-based Fine-tuning

For the downstream task, the constructed prompt function would be set as input to the language model M and make the prediction for its sentiment label. In this step, we combine all semantic descriptions for each modality with the static template of *The man is saying "T_t" with T_v and T_a*. Adding with the text template *It is* as a prefix prompt. The complete multimodal prompt P_m for input should be:

$$P_m^i = It\ is\ [Mask].T_c^i \tag{3.5}$$

The incorporation of the multimodal prompt P_m^i proves pivotal in the fine-tuning stage when introduced to the PLM M. The BERT[37] serves as the foundational architecture for our approach, and we execute prompt-based fine-tuning. The determination of the optimal candidate, based on superior performance on the dev dataset, guides the construction of the prompt within the generated label mapping space. After feeding the prompt P_m^i into M, the hidden vectors would be computed as shown in 3.2, and probability is calculated according to specific word tokens from label mapping space V. The formula of probability is denoted as:

$$p(y|x_{in}) = p([Mask] = M(y)|P_m) = \frac{exp(\mathbf{w}_{M(y)*\mathbf{h}_{[Mask]}})}{\Sigma_{y' \in y} exp(\mathbf{w}_{M(y')*\mathbf{h}_{[Mask]}})}, \tag{3.6}$$

where $h_{[Mask]}$ is the hidden vector of $[Mask]$ and $\mathbf{w}$ denotes the pre-softmax vector, M indicates the mapping rule from label space to specific words in the LM. M can be fine-tuned to minimize the cross-entropy loss.

The classification mechanism is deeply rooted in BERT. In this process, the slot in the constructed prompt is replaced with the corresponding word from the selected label space. Following this substitution, the entire input is presented to the PLM, yielding the prediction for the label. Figure 3.4

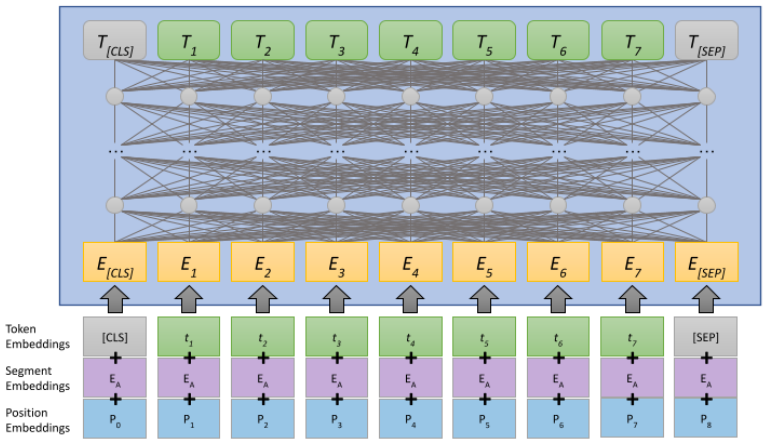

Figure 3.2: The representation generated by BERT

provides an illustrative representation of this stage. The representations of the input, augmented with the masked token, undergo processing through the 24 encoders with the detailed structure shown in Figure. 3.3, and apply with the loss function of CrossEntropy, defined as:

$$-\sum_{c=1}^{3} y_{x_i,l} \log(p_{x_i,l}) \tag{3.7}$$

where x_i denotes the i_{th} sample and l represents the label, and the optimizer of AdamW[38].

Besides, our chosen label space is based on the search from PLM automatically, but also several studies have highlighted that both manual prompts and auto prompts possess their own merits. In general, manual prompts are informed by expert knowledge within a specific research domain and are highly regarded for their interpretability from the perspective of natural language. While auto prompts are automatically generated by LMs with seemingly a lack of explication in natural language. Thus, we also test the manual prompt result in our experiments.

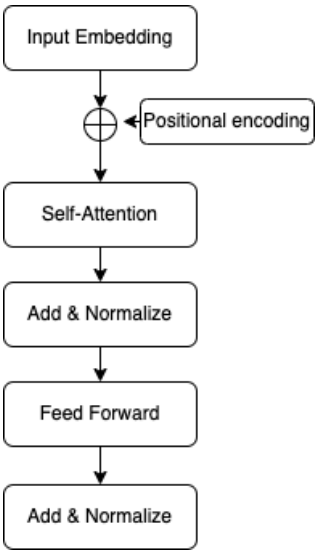

Figure 3.3: Encoder structure in BERT

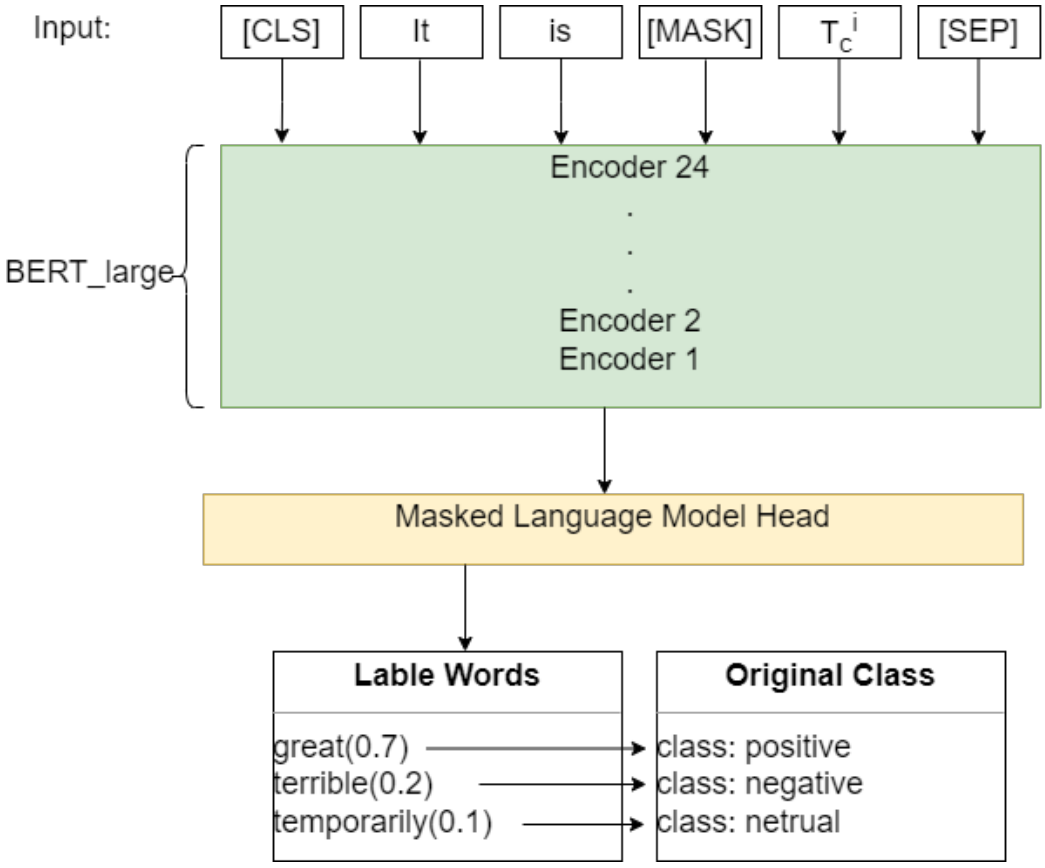

Figure 3.4: Fine-tuning based on prompt in BERT$_{LARGE}$

Chapter 4

Experiments

4.1 Dataset

The Multimodal EmotionLines Dataset(MELD)[21] is widely recognized and frequently employed for MSA, featuring annotations for both sentiment and emotion. In our experiments, we exclusively focus on utilizing sentiment annotations, specifically categorizing instances into positive, negative, and neutral labels. The samples from the dataset comprise a rich combination of audio, visual, and textual modalities, incorporating content from over 1400 dialogues and encompassing more than 13,000 utterances extracted from the popular "Friends" TV series. A comprehensive statistical summary detailing the distribution of the dataset across different modalities and sentiments is provided in Table 4.1. This table encapsulates essential information regarding the dataset split.

	#utterance		
Sentiment	Train	Dev	Test
positive	2334	233	521
negative	2945	406	833
neutral	4710	470	1256
	9989	1109	2610

Table 4.1: Data split on MELD

4.2 Experimental Setup

At first, since BERT[37], is a PLM renowned for its robust capabilities, we choose it as the foundational architecture for our experimental framework.

Specifically, we build upon the BERT$_{LARGE}$ variant to harness its extensive pre-trained knowledge. The figurations of BERT$_{LARGE}$ are: 24 layers, 1024 hidden dimensions, 16 attention heads, and 340M parameters. Our proposed method involves fine-tuning the model on the training dataset sourced from the MELD corpus.

In terms of hyperparameter settings, we adopt a thoughtful approach to ensure optimal training. The number of epochs is set to 50, providing a balance between convergence and computational efficiency. Additionally, the learning rate is configured at 1e-5, a common choice that facilitates steady convergence during the fine-tuning process.

To ensure the robustness and reliability of our results, we conduct multiple experiments using three distinct random seeds, e.g. 16, 30, and 56. This repetition allows us to account for variations introduced by different initializations. The final reported results are derived from the average performance across these experiments, providing a comprehensive assessment of the model's efficacy under diverse conditions on the test dataset. Our chosen evaluation metric is the averaged weighted F1 score (F1), a well-established measure for assessing the precision and recall of our model. The following formulas define key metrics:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \tag{4.1}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4.2}$$

Where "TP" is the number of true positives, "FN" is the number of false negatives, and "FP" is the number of false positives.

$$\text{Averaged Weighted F1} = \frac{\sum_i W_i \times F1_i}{n}, \tag{4.3}$$

where i denotes the i_{th} class, W_i represents the percentage of the number of class i in all classes, and n represents the total number of class labels.

This metric takes into account both false positives and false negatives, offering a comprehensive evaluation of the model's performance.

4.3 Baseline

We conduct a comparative analysis between our proposed method and the original baseline outlined in [21]. Recognizing the temporal aspect of the published paper, which may not capture recent advancements, we introduce

an additional baseline for an updated evaluation. This baseline is called as Ubaseline in the following section. Since the baseline from original of MELD, published in 2019. The baselines applied with LSTM or RNN model architecture, are not fair to be compared with recent enhanced attention mechanism based on PLM. We set an supplementary baseline to make a fair comparison to our proposed method. For this supplementary baseline, we adopt a fundamental approach in MSA. Modalities are individually processed, with the following feature extractors employed: SentenceTransformer[39] for text, OpenFace[33] and MANet[40] for visual data, and OpenSmile[35] using the IS09 feature set for audio. The visual and audio feature extractor are the same as what we employ in our proposed method to keep a balance on comparison. The text feature extractor applied with a enhanced tranformer-based extractor. At the fusion stage, multimodal features are generated through the concatenation of embeddings produced by each extractor directly. Subsequently, these multimodal features are fed into an attention mechanism classifier for evaluation. This comparison enables a comprehensive assessment of the proposed method against both the original baseline and a contemporary basic MSA approach, facilitating a thorough understanding of its performance across different experimental settings.

4.4 Result

The performance comparison between our proposed model and the baseline is presented in Table 4.2. Unless explicitly specified, the designation 'A' signifies textual information derived from audio features, 'V' denotes textual information extracted from visual features, and 'T' represents textual information derived from utterances in all subsequent tables. In [21], the baselines outlined are bcLSTM[41] and DialogueRNN[42]. However, it is important to note that these baselines, while effective for their time, exclusively handle audio and text modalities, as the management of visual modality presented challenges during that period.

Model	Mode	F1-score
bcLSTM	T+A	66.68
DialogueRNN	A+T	67.56
Ubaseline	T+V+A	61.58
Proposed model	T+V+A	71.17

Table 4.2: Test set average weighted F1-score results on MELD

The bcLSTM model generates representations by employing a bi-directional

Recurrent Neural Network (RNN) with a hierarchical process. This approach initially models unimodal context and subsequently incorporates bi-modal context features. DialogueRNN adopts a model architecture that utilizes Gated Recurrent Units (GRU) across three stages to capture emotional context in conversations. Utterances undergo processing through both global and part GRUs to update states, and an emotion GRU is employed to model emotional information for classification based on the updated state by the other two GRUs. These baseline models serve as benchmarks for our proposed model's performance, allowing for a comprehensive evaluation across different modalities and contextual considerations.

In evaluating the outcomes of our proposed method, it is imperative to articulate a precise statement outlining the procedures employed. During the label search phase, we meticulously select ten candidate groups based on their efficacy in mapping label words to alternative terms, thereby demonstrating superior performance on the development dataset ranking. Subsequently, we identify the group that attains the highest level of performance. This selected group is then utilized to substitute the [Mask] slot in our input when interfacing with BERT. The specific words constituting this optimal group are as follows: "neutral": "stitches", "positive": "Automatic", "negative": "portfolio". This group means that each senimental ananotation in dataset is replaced to its corresponding word in this word group. This is also the group we utilized for our main result, as shown in above table.

Examining these terms from the standpoint of natural language may raise concerns about apparent dissimilarities. To address this, we acknowledge the need for an in-depth discussion within the realms of the Discussion section. In this subsequent discourse, we aim to delve into the intricacies of this issue, presenting additional experimental results derived from diverse mapping rules. These findings will contribute to a nuanced understanding of the observed semantic variations and inform potential refinements to our methodology.

The empirical findings of our study unequivocally demonstrate the superior performance of our proposed method when juxtaposed with the results obtained from baseline models. Specifically, our method exhibits a remarkable improvement, surpassing the performance of established baselines.

Notably, our approach outperforms bcLSTM by a margin of 5 percent, DialogueRNN by 3 percent, and Ubaseline by 10 percent, respectively. The discernible effectiveness of our proposed method can be attributed to the innovative integration of the prompt-based approach. This methodology proves instrumental in mitigating performance gaps between different modalities. In particular, our approach focuses on generating semantic descriptions, thereby leveraging the latent power embedded within Large Language Models (LLM).

This strategic incorporation of prompt-based techniques contributes significantly to the heightened performance observed across various metrics.

The key to our method's success lies in its ability to bridge the disparities inherent in multimodal sentiment analysis. By employing prompt-based techniques, we alleviate the challenges associated with diverse modalities and enhance the model's capacity to discern intricate nuances within the data. The generation of semantic descriptions not only serves as a unifying bridge but also facilitates a more comprehensive understanding of sentiment expressions across different modalities.

Furthermore, the utilization of Large Language Models adds an additional layer of sophistication to our method. The inherent capabilities of LLM, including contextual understanding and semantic richness, synergize with the prompt-based approach to yield a holistic and effective solution. The synergy between these components is pivotal in achieving the observed performance enhancement, providing a promising avenue for future research in multimodal sentiment analysis.

In essence, our proposed method stands as a testament to the efficacy of prompt-based techniques in addressing the challenges posed by diverse modalities. The notable improvements over baseline models underscore the significance of our method in advancing the state-of-the-art in multimodal sentiment analysis. Looking forward, our research paves the way for further exploration and refinement of prompt-based methodologies in the dynamic landscape of multimodal analysis.

4.5 Ablation Analysis

We conduct ablation studies for each modality, and the results are presented in Table 4.3. It is important to note that our proposed method primarily handles visual and acoustic modalities in a textual manner, thus all results are based on text alone. The findings from the unimodal results indicate that using solely semantic descriptions from audio features yields the lowest evaluation, while textual information achieves the highest evaluation. This aligns with our expectations, as our proposed method is built upon text, and text modality inherently contains the most information compared to other modalities. The comparatively lower performance of audio can be attributed to our extraction of audio features limited to three types, including only numerical data. Additionally, PLM exhibit reduced sensitivity when handling data with numerical types, contributing to the observed results. In the bimodal analysis, the combination of textual (T) and visual (V) modalities achieves the highest evaluation. This result suggests that, in the context of

this research, these two modalities contribute more significantly compared to the audio feature.

Mode	F1-score
T	61.68
V	57.56
A	54.05
T+A	62.47
T+V	63.03
A+V	60.78
T+V+A	71.17

Table 4.3: Results on combination of unimodal and bimodal

Moreover, to analyze the contribution from each phase in our proposed method, we conduct an ablation study in this section. We divide our proposed method into two main phases. The first phase involves using a verbalizer to construct a label mapping space with assistance from PLM. The second phase entails feeding the completed prompt function, along with the extracted data, into the PLM to make predictions. These phases are correspondingly referred to as P1 and P2 in the following statements. The experimental results are presented in Table 4.4. From these results, we observe that if we fine-tune our model without the verbalizer (P1), the performance significantly decreases. At P1, we solely apply with the original label word as verbalizer with exact label mapping. This finding underscores the essential role of the verbalizer in the prompt-based fine-tuning method.

Phase	Mode	F1-score
P2	T+A+V	63.68
P1+P2	T+A+V	71.17

Table 4.4: Results on phases

Chapter 5

Discussion

This section extends the discussion by presenting additional experimental results, providing insights into how our model performs under varied settings. In our primary results, we emphasize the utilization of the Best word mapping. However, in this section, we broaden the scope by showcasing experimental outcomes for multiple label mapping groups, as illustrated in Table 5.1. All the results are obtained under the same experimental setting as the main results, ensuring consistency. Furthermore, the reported results in the following table are based on averaging across three distinct random seeds to enhance robustness.

To elucidate the specificities of each label mapping group, Table 5.2 presents the extracted words corresponding to each group. This comprehensive overview not only enriches our understanding of the experimental outcomes but also facilitates a comparative analysis of the performance across diverse label mappings.

Label Mapping	Mode	F1-score
Mapping_1	T+A+V	71.17
Mapping_2	T+A+V	64.68
Mapping_3	T+A+V	63.35
Mapping_4	T+A+V	63.47
Mapping_5	T+A+V	67.27

Table 5.1: Test result on different label mapping group

From the result for different label mapping group as shown in Table 5.1, We observe that different groups exhibit varying impacts on the predictive outcomes, with certain groups demonstrating differences of up to 4 percentage points. Notably, the first group corresponds to the set employed in our primary experimental results. The highest value is evident in the first

Label Mapping	Words Group
Mapping_1	"neutral": "stitches", "positive": "Automatic", "negative": "portfolio"
Mapping_2	"negative": "protested", "neutral": "investigator", "positive": "concede"
Mapping_3	"negative": "aspire", "neutral": "investigator", "positive": "frowned"
Mapping_4	"negative": "ingest", "neutral": "investigator", "positive": "absorb"
Mapping_5	"negative": "protested", "neutral": "temporarily", "positive": "pH"

Table 5.2: Label mapping groups

group, with a specific F1 score exceeding 71, while the lowest value is associated with the third group with the performance of 63.35. Respectively, the label mapping groups corresponding to:"neutral": "stitches", "positive": "Automatic", "negative": "portfolio" and "negative": "aspire", "neutral": "investigator", "positive": "frowned".

From the perspective of natural language semantics, there appears to be limited similarity between the label words and their corresponding terms. However, during the generation of verbalizers by Pre-trained Language Models (PLM), these results emerge. We posit that this outcome may be attributed to the inherent contextual coherence of these words in the original training corpus of BERT. In other words, words with semantic proximity to the original label terms within the semantic space of the training corpus tend to yield superior classification results when employed in the method proposed in this study.

In the course of this research, certain limitations have surfaced, warranting careful consideration within specific domains. Firstly, an inherent challenge lies in the labels generated by the language model. It is observed that certain terms lack semantic coherence when examined from the standpoint of natural language. More precisely, these terms do not align closely with the definitions encapsulated by the designated label words. Remarkably, despite this semantic incongruity, the model manages to yield commendable results. The interpretability of the model, however, becomes a focal point for improvement. Addressing this issue is imperative for enhancing the transparency of the model's decision-making process. Subsequent investigations in future research endeavors are anticipated to delve deeper into unraveling the intricacies of this particular challenge.

Another pertinent limitation pertains to the static nature of the templates employed throughout our experimental design. The fixed template structure constrains the breadth of our discourse on the prompt-based methodology. Future research initiatives are envisioned to involve a more dynamic exploration of varied templates. This methodological refinement aims to unlock

a richer understanding of the nuances associated with the prompt-based approach, thereby contributing to a more comprehensive and nuanced interpretation of the method's efficacy.

Chapter 6

Conclusion

In summary, we introduced a text prompt-based approach for addressing the challenges posed by multimodal sentiment analysis tasks. Our experiments, conducted on the MELD dataset, showcase the effectiveness of this method. This innovative approach capitalizes on the substantial advancements in pretrained language models, tapping into a rich source of information. The utilization of a prompt-based strategy not only facilitates a more interpretable and streamlined model architecture but also serves to alleviate disparities between different modalities, expediting the fine-tuning step in multimodal sentiment classification tasks.

Particularly noteworthy is the demonstrated interpretability in handling various modalities and orchestrating their combination. Despite the manual generation of most combination rules, the final results exhibit a commendable level of performance at this stage. Furthermore, our findings strongly suggest untapped potential within pretrained language models.

Lastly, an overarching aspiration involves expanding our understanding of the prompt-based fine-tuning method beyond the confines of the current research scope. This aspiration stems from the recognition that the method may harbor untapped potential in diverse research domains. Unraveling and harnessing this latent potential requires future investigations to extend beyond the immediate scope of this research. The prospect of discovering novel applications and refining the method's adaptability in hitherto unexplored fields remains a focal point for subsequent research endeavors.

In summation, while this research has yielded valuable insights, the identified limitations underscore the need for ongoing refinement and expansion. Addressing these limitations will not only bolster the internal validity of our findings but also serve as a catalyst for future explorations in the evolving landscape of prompt-based fine-tuning methodologies.

References

[1] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," in *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*. Springer, 2016, pp. 15–27.

[2] D. Cao, R. Ji, D. Lin, and S. Li, "A cross-media public sentiment analysis system for microblog," *Multimedia Systems*, vol. 22, pp. 479–486, 2016.

[3] Z. Xu, V. Pérez-Rosas, and R. Mihalcea, "Inferring social media users' mental health status from multimodal information," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6292–6299.

[4] R. Walambe, P. Nayak, A. Bhardwaj, and K. Kotecha, "Employing multimodal machine learning for stress detection," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–12, 2021.

[5] Y. Li, S. Wang, Q. Pan, H. Peng, T. Yang, and E. Cambria, "Learning binary codes with neural collaborative filtering for efficient recommendation systems," *Knowledge-Based Systems*, vol. 172, pp. 64–75, 2019.

[6] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[7] G. Cai and B. Xia, "Convolutional neural networks for multimedia sentiment analysis," in *Natural Language Processing and Chinese Computing: 4th CCF Conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings 4*. Springer, 2015, pp. 159–167.

[8] P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," *arXiv preprint arXiv:1907.01011*, 2019.

[9] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger, "Explainable ai: the new 42?" in *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2*. Springer, 2018, pp. 295–303.

[10] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Generating counterfactual explanations with natural language," *arXiv preprint arXiv:1806.09809*, 2018.

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

[12] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[14] J. Zhao, R. Li, Q. Jin, X. Wang, and H. Li, "Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4703–4707.

[15] H. Wu and X. Shi, "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)*, 2022, pp. 2438–2447.

[16] D. Hazarika, Y. Li, B. Cheng, S. Zhao, R. Zimmermann, and S. Poria, "Analyzing modality robustness in multimodal sentiment analysis," *arXiv preprint arXiv:2205.15465*, 2022.

[17] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 481–492.

[18] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional lstm," *Multimedia Tools and Applications*, vol. 80, pp. 13 059–13 076, 2021.

[19] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[20] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[21] P. Soujanya, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "A multimodal multi-party dataset for emotion recognition in conversations," 2018.

[22] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.

[23] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," *arXiv preprint arXiv:2012.15723*, 2020.

[24] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," *arXiv preprint arXiv:2010.15980*, 2020.

[25] E. M. Ponti, G. Glavaš, O. Majewska, Q. Liu, I. Vulić, and A. Korhonen, "Xcopa: A multilingual dataset for causal commonsense reasoning," *arXiv preprint arXiv:2005.00333*, 2020.

[26] R. Shin, C. H. Lin, S. Thomson, C. Chen, S. Roy, E. A. Platanios, A. Pauls, D. Klein, J. Eisner, and B. Van Durme, "Constrained language models yield few-shot semantic parsers," *arXiv preprint arXiv:2104.08768*, 2021.

[27] T. Schick and H. Schütze, "Exploiting cloze questions for few shot text classification and natural language inference," *arXiv preprint arXiv:2001.07676*, 2020.

[28] X. Yang, S. Feng, D. Wang, Y. Zhang, and S. Poria, "Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6045–6053.

[29] A. Brock, S. De, and S. L. Smith, "Characterizing signal propagation to close the performance gap in unnormalized resnets," *arXiv preprint arXiv:2101.08692*, 2021.

[30] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[31] Y. Yu and D. Zhang, "Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[34] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[36] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[39] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[40] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[41] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[42] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.