Mixture-of-Prompt-Experts for Multi-modal Semantic Understanding

Zichen Wu, Hsiu-Yuan Huang, Fanyi Qu and Yunfang Wu*
National Key Laboratory for Multimedia Information Processing, Peking University, China
School of Computer Science, Peking University, China
wuzichen@pku.edu.cn, wuyf@pku.edu.cn

Abstract

Deep multi-modal semantic understanding that goes beyond the mere superficial content relation mining has received increasing attention in the realm of artificial intelligence. The challenges of collecting and annotating high-quality multi-modal data have underscored the significance of few-shot learning. In this paper, we focus on two critical tasks under this context: few-shot multi-modal sarcasm detection (MSD) and multi-modal sentiment analysis (MSA). To address them, we propose Mixture-of-Prompt-Experts with Block-Aware Prompt Fusion (MoPE-BAF), a novel multi-modal soft prompt framework based on the unified vision-language model (VLM). Specifically, we design three soft prompt experts: a text prompt and an image prompt that extract modality-specific features to enrich the single-modal representation and a unified prompt to assist multi-modal interaction. Additionally, we reorganize Transformer layers into several blocks and introduce cross-modal prompt attention between adjacent blocks, which smoothens the transition from single-modal representation to multi-modal fusion. On both MSD and MSA datasets in few-shot settings, our proposed model not only surpasses the 8.2B model InstructBLIP with merely 2% parameters (150M), but also significantly outperforms other widely-used prompt methods on VLMs or task-specific methods.

Keywords: multi-modal sarcasm detection, multi-modal sentiment analysis, prompt learning

1. Introduction

Multi-modal semantic understanding (MSU) is crucial for the development of machines capable of interpreting the complex interplay of textual and visual information. In social media platforms, where the combination of text and imagery can often present conflicting messages or nuanced sentiments that are not immediately apparent from a single modality alone, such understanding is vital for accurately interpreting the intent and sentiment. Among the fields of MSU, Multi-modal Sarcasm Detection (MSD) and Multi-modal Sentiment Analysis (MSA) emerge as two representing tasks. These tasks exemplify the intricate process of aligning and comprehending the relations from different modalities to discern the intended meaning or sentiment. As highlighted by the examples in Table 1, solely reading the text or image of MSD is prone to misinterpreting it as a positive comment while ignoring the sarcasm about the discount being too small.

Recent approaches for MSU normally exploit a dual-encoder architecture, i.e., use separate pre-trained encoders to extract features for different modalities (e.g., BERT (Devlin et al., 2019) for text and ResNet (He et al., 2016) for image). The features then interact with each other to capture the incongruity and fed into a classification head for the prediction. Under this framework, researchers are dedicated to designing effective methods of interaction, including attention mechanisms (Pan et al., 2020; Xu et al., 2020; Han et al., 2021a), graph structures (Liang et al., 2021, 2022; Liu et al., 2022a), optimal transport (Pramanick et al., 2022) and dynamic routing (Tian et al., 2023).

Task	MSD	MSA
Image		
Text	Great Sale!	Wish you a good Valentines Day.

Table 1: Examples of MSD and MSA task. The left example displays *sarcasm*, as the text "great sale" contradicts with the image depicting a mere 1-dollar discount. The right example conveys a positive wishing attitude, thus categorized as *positive*.

Although the previous studies have achieved good performance in semantic understanding tasks like MSD or MSA, they mostly rely on sufficient training data. However, collecting a large amount of high-quality multi-modal data, sarcasm especially, is a non-trivial task. According to Misra and Arora (2023), sarcasm expressions require a high level of cognitive ability and rarely appear on social media, making it difficult to collect and annotate. Moreover, most existing models use separate pre-trained encoders to process text and image, which might lead to the misalignment of different modalities and thus hurt modal fusion. Nowadays, pre-trained on large-scale image-text pairs, the vision-language models (VLMs) achieve good image-text correspondences and can perform well on cross-modal reasoning. Given these two aspects, we propose to address

* Corresponding author.

LREC-COLING 2024, pages 11381–11393

the few-shot MSU tasks by using VLMs.

To adapt the pre-trained VLMs to downstream multi-modal tasks on few-shot settings, prompt-based learning is widely applied and has demonstrated promising performance (Liu et al., 2023a). Compared to manually designed prompts, continuous prompts (also referred to as soft prompts) are preferred due to their flexibility and scalability (Liu et al., 2021a,b; Han et al., 2021b). However, most previous work only coped with text data. In this paper, we explore different methods of utilizing soft prompts to address the few-shot MSU task.

Two primary architectures are prevalent in VLMs. One line of work encodes images and texts respectively and performs modal fusion by simply computing the similarities between them, like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). Another line of work adopts a unified network for both single-modal representation and multi-modal fusion, like VilBert (Lu et al., 2019) and VLMo (Bao et al., 2022). In the first line of work, the soft prompt method has been studied, like CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), UPT (Zang et al., 2022) and MaPLe (Khattak et al., 2023). Nonetheless, there has been little work applying soft prompts in the second line of VLMs yet. Thus, exploring multi-modal prompts in a unified Transformer network remains an open issue.

In this paper, towards the deep MSU, we propose a novel multi-modal soft prompt framework MoPE-BAF, **M**ixture-**o**f-**P**rompt-**E**xperts with **B**lock-**A**ware prompt **F**usion. Specifically, we devise a set of soft prompts corresponding with different roles, an image prompt expert, a text prompt expert and a unified prompt expert. The first two extract semantic features within a single modality, while the third assists in capturing inter-modality information. Furthermore, we introduce a block-aware prompt fusion mechanism to enhance the connection between different prompt experts. We re-organize the transformer layers into several blocks and apply cross-attention to enable the exchange of prompt expert information between two adjacent blocks. It facilitates deep interactions between modalities and enables smoother transitions from single-modal representation to multi-modal fusion.

We conduct a series of experiments on the MSDT dataset (Cai et al., 2019) in the few-shot setting. Our proposed model significantly outperforms the classical CLIP and the base VLMo. More importantly, our model with only 150M parameters obtains a better performance than InstructBLIP (Dai et al., 2023), the most advanced model with 8.2B parameters. Besides, experimental results demonstrate a stable performance gain compared to conventional soft prompts, no matter using a LM head or a classification head. Furthermore, we apply our model to the MSA task on the MVSA-S data (Niu et al., 2016), outperforming the previous state-of-the-art UP-MPF (Yu et al., 2022) by 3.91 F1 points.

To sum up, our contributions are:

- We propose a novel mixture-of-prompt-experts method on the unified VLMs, which drives the pre-trained model to get refined in both single-modal representation and multi-modal fusion.
- We present a block-aware prompt fusion mechanism, which activates deep interactions between prompt experts and balances the twin objectives of single-modal specialization and multi-modal fusion in VLMs.
- We conduct experiments on the few-shot multi-modal sarcasm detection and sentiment analysis, outperforming the previous state-of-the-art methods and the advanced large language models.

2. Related Work

2.1. Multi-modal Sarcasm Detection

Researchers have been exploring different methods to model the incongruity within the image-text pair for the MSD task. At the outset, feature-based approaches are adopted (Schifanella et al., 2016; Castro et al., 2019). Subsequently, researchers pay more attention to modality interaction. Cai et al. (2019) leverages a hierarchical strategy to fuse the three modalities of image, attribute and text by attention weights. Sangwan et al. (2020) exploits the interaction among the input modalities using the recurrent neural network. Pan et al. (2020) tries to capture the incongruity by inter-modal attention and co-attention within the text. Xu et al. (2020); Han et al. (2021a) decompose the model into a separation network for discrepancy increment and a relation network for relevance increment. (Liang et al., 2021, 2022; Liu et al., 2022a) introduce graph structures for depicting incongruity relations. More recently, Tian et al. (2023) utilizes dynamic paths to activate different routing Transformers. However, these works rely on a large amount of training data to finetune the pre-trained models.

2.2. Multi-modal Sentiment Analysis

In recent years, MSA has become a popular research topic. Xu et al. (2018) designs a co-memory network based on attention to predict the whole sentiment of text-image pairs. Yang et al. (2021a) proposes a multi-view attention network, which utilizes an attention memory network to extract text and image features, and then fuses multi-modal features through a stacking-pooling module. After that, they incorporated graph neural network in MSA (2021b). Yu et al. (2022) proposes a unified

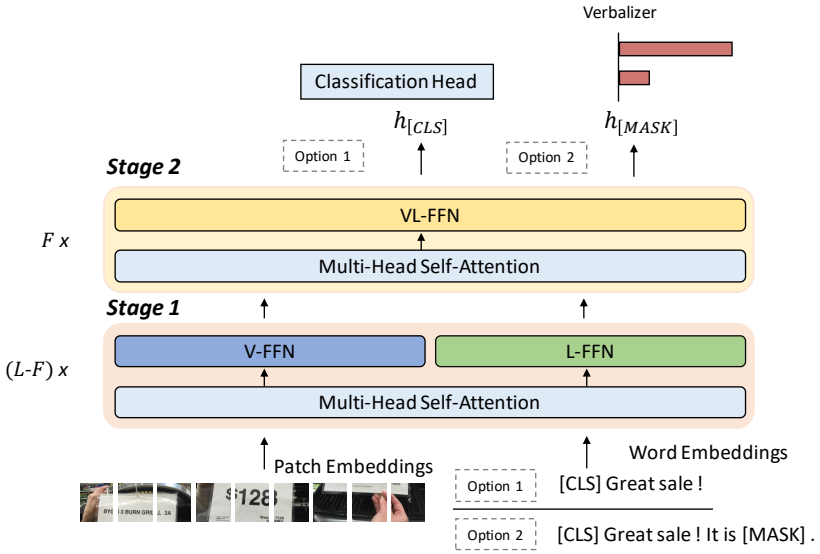

Figure 1: The framework of VLMo for image-text detection task. We demonstrate two methods. In *finetuning*, the [CLS] representation is fed to a classification head, while in *manual prompt*, the representation of [MASK] is fed to a verbalizer.

pre-training stage to narrow the semantic gap between different image and text pre-trained models.

2.3. Multi-modal Prompt Learning

Recently, researchers adapt VLMs to fit downstream tasks via prompt learning (Gu et al., 2023), which can be generally categorized into single-modal prompt and multi-modal prompt methods. In single-modal prompt methods, learnable continuous prompts are appended in front of the text data. For example, CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) replace the manual-crafted text prompts used in CLIP with continuous vectors for image classification. Multi-modal prompt learning aims to optimize the text and image inputs simultaneously. UPT (Zang et al., 2022) designs a shared initial prompt for CLIP text and visual encoders. MaPLe (Khattak et al., 2023) leverages prefix tuning in both modality encoders and designs a coupling function to enable a mutual promotion between prompts. CMPA (Liu et al., 2023b) designs prompts for both encoders and employ a cross-modal prompt attention at each layer. These prompt-based methods are mainly applied to the CLIP architecture, which encodes the text and image inputs separately. To the best of our knowledge, there has been little work that applies soft prompts to a unified vision-language pre-trained model.

3. Preliminary

Since our work adopts the vision-language pre-trained model VLMo as the backbone, to better illustrate our proposed method, we provide an overview of VLMo as well as the basic knowledge of applying it to MSU tasks.

3.1. The Vision-language Pre-trained Model: VLMo

VLMo presents a Transformer architecture with Mixture-of-Modality-Experts where the feed-forward neural (FFN) network switches based on the input modality and fusion requirements. As shown in Figure 1, given an image-text pair, VLMo performs the unified encoding in two stages: (1) Employing vision FFN (V-FFN) and language FFN (L-FFN) to encode the respective modality representations at the bottom Transformer layers, and (2) Using vision-language FFN (VL-FFN) to perform multi-modal interaction at the top layers.

Concretely, in Stage 1, denoting the hidden vision and language representations of the previous layer as H_v^{n-1}, H_l^{n-1} respectively, VLMo first employs shared self-attention across modalities to align their contents, and then pass them to a uni-modal FFN to obtain the output of this layer:

$$H^{n-1} = \text{concat}(H_v^{n-1}, H_l^{n-1}), \quad (1)$$

$$\tilde{H}^n = \text{softmax}(\frac{(H^{n-1}W_q)(H^{n-1}W_k)^\top}{\sqrt{d}})(H^{n-1}W_v), \quad (2)$$

$$[\tilde{H}_v^n, \tilde{H}_l^n] = \tilde{H}^n \quad (3)$$

$$H_v^n = \text{V-FFN}(\tilde{H}_v^n), \quad (4)$$

$$H_l^n = \text{L-FFN}(\tilde{H}_l^n) \quad (5)$$

In Stage 2, after the self-attention operation, the intermediate outputs are combined and forwarded to a vision-language FFN:

$$H_v^n, H_l^n = \text{VL-FFN}([\tilde{H}_v^n, \tilde{H}_l^n]). \quad (6)$$

3.2. Tuning on Multi-modal Tasks

In MSD or MSA task, the input consists of an image x and associated text y, with the goal of predicting its corresponding category.

Finetuning. The most straightforward approach is to register a classification head on top of VLMo. After encoding, the vector of the text-start token ([CLS]) is used as the final representation of the image-text pair, and the prediction result is obtained:

$$p_t = \mathbb{P}(t|\theta, \phi, x, y), \tag{7}$$

where θ, ϕ denote the parameters of the pre-trained model and the classification head respectively.

Manual Prompt. Another way to utilize VLMo is to reformulate the task into a mask language modeling task using a hard template containing [MASK]:

$$p_t = \mathbb{P}([\text{MASK}] = v(t)|\theta, T(x, y)), \tag{8}$$

where T denotes the prompt template and v denotes the verbalizer that maps the [MASK] token to the probabilities on label words.

Soft Prompt. The soft prompt method utilizes a set of trainable virtual tokens, which eliminates the need to manually design templates:

$$T(x, y) = [V_1, V_2, \cdots, V_n]\ x\ ||\ y. \tag{9}$$

where $\{V_i\}$ is the virtual token.

4. The Proposed Model

4.1. Mixture-of-Prompt-Experts

One of the primary challenges in applying soft prompts to multi-modal tasks is the specialization of prompt properties across different modalities. In the traditional soft prompt method, all prompts and input tokens are treated equally within Transformer layers, which aligns with the VLMo's second stage. However, VLMo differentiates image and text inputs with distinct FFNs in the first stage. Thus, only the second stage can be activated by prompts, which inhibits the full exploitation of the multi-modal encoder in extracting modality-specific features.

In view of this, we propose a novel multi-modal soft prompt approach, Mixture-of-Prompt-Experts (MoPE), to serve the two encoding stages in VLMo. It contains a set of soft prompts corresponding to the three functional FFNs of VLMo: image expert (V-Prompt), text expert (L-Prompt) and unified cross-modal expert (VL-Prompt), where V-Prompt and L-Prompt assist in extracting the semantic features from the respective modality in Stage 1, and VL-Prompt assists in enhancing inter-modal interaction. Each prompt expert is initialized as a set of trainable vectors with dimension matching the embeddings of the pretrained model. The overall structure of our proposed model is illustrated in Figure 2.

Stage 1. We prepend a single-modal prompt to each modality input and pass them together through VLMo Transformer layers.

$$H^{n-1} = \text{concat}(H^{n-1}_{vp}, H^{n-1}_{lp}, H^{n-1}_{v}, H^{n-1}_{l}), \tag{10}$$

$$\tilde{H}^{n} = \text{softmax}(\frac{(H^{n-1}W_q)(H^{n-1}W_k)^\top}{\sqrt{d}})(H^{n-1}W_v), \tag{11}$$

$$[\tilde{H}^{n}_{vp}, \tilde{H}^{n}_{lp}, \tilde{H}^{n}_{v}, \tilde{H}^{n}_{l}] = \tilde{H}^{n} \tag{12}$$

$$[H^{n}_{vp}, H^{n}_{v}] = \text{V-FFN}([\tilde{H}^{n}_{vp}, \tilde{H}^{n}_{v}]), \tag{13}$$

$$[H^{n}_{lp}, H^{n}_{l}] = \text{L-FFN}([\tilde{H}^{n}_{lp}, \tilde{H}^{n}_{l}]) \tag{14}$$

where H_v, H_l, H_{vp}, H_{lp} denote vision representation, language representation, V-Prompt and L-prompt, respectively.

To ensure the specialization of prompt experts within their corresponding modalities, we restrict their receptive field in the self-attention module, as illustrated in Figure 3. V-Prompt is dedicated to the image input only, while the image can attend to both V-Prompt and the text input, enjoying the cross-modal alignment and uni-modal enhancement simultaneously. L-Prompt performs in the same way.

Stage 2. A multi-modal unified prompt expert is introduced to enhance the interaction between modalities:

$$H^{n-1} = \text{concat}(H^{n-1}_{vlp}, H^{n-1}_{v}, H^{n-1}_{l}), \tag{15}$$

$$\tilde{H}^{n} = \text{softmax}(\frac{(H^{n-1}W_q)(H^{n-1}W_k)^\top}{\sqrt{d}})(H^{n-1}W_v), \tag{16}$$

$$[\tilde{H}^{n}_{vlp}, \tilde{H}^{n}_{v}, \tilde{H}^{n}_{l}] = \tilde{H}^{n}, \tag{17}$$

$$[H^{n}_{vlp}, H^{n}_{v}, H^{n}_{l}] = \text{VL-FFN}([\tilde{H}^{n}_{vlp}, \tilde{H}^{n}_{v}, \tilde{H}^{n}_{l}]), \tag{18}$$

where H_{vlp} is the representation of VL-Prompt.

4.2. Block-Aware Prompt Fusion

Recall that deep MSU necessitates the ability to discern the complex relationships across image and text modalities. These tasks demand a profound understanding of content relations beyond simple fusion. Despite VLMs being pre-trained on large image-text corpora and demonstrating strong performance on traditional multi-modal tasks (e.g., image classification), their performance on the complex multi-modal task remains subpar due to the superficial fusion process. In VLMo, the allocation of fusion layers is limited (e.g., VLMo-Base-plus assigns only 3 layers), constraining the model's fusion capacity. Accordingly, we propose a new **b**lock-**a**ware prompt **f**usion (BAF) mechanism to

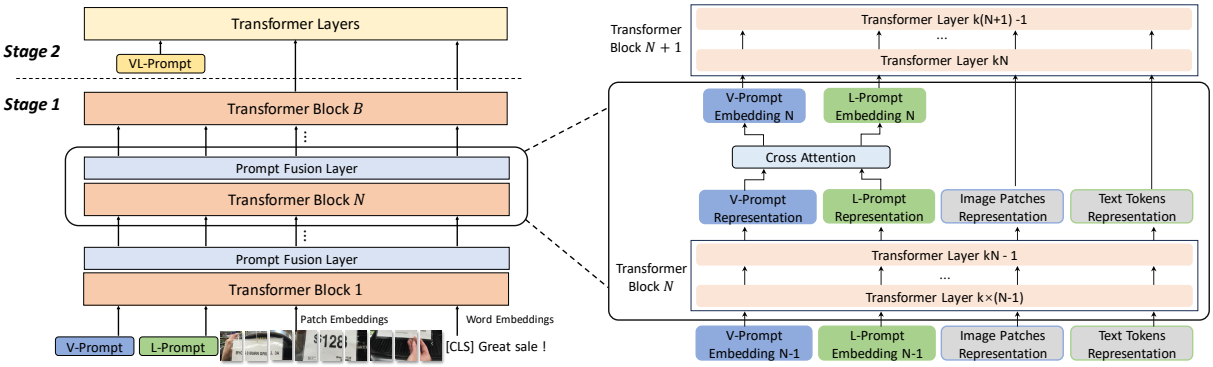

Figure 2: Our proposed MoPE-BAF model for multi-modal semantic understanding .

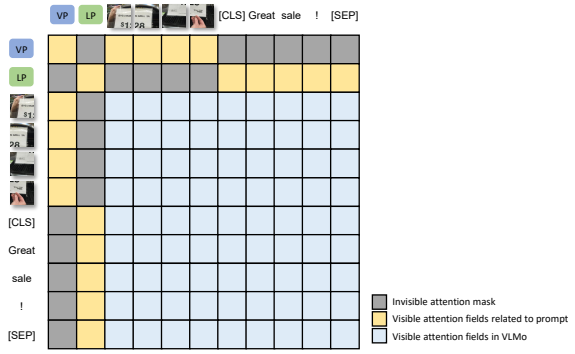

Figure 3: Receptive fields of different prompts, image patches, text tokens in the self-attention module when using MoPE. VP and LP are shorthand for V-Prompt, L-Prompt.

make different modality prompts interact deeply and meet the fusion requirements in deep MSU.

To be specific, we re-organize the transformer layers into several blocks, and introduce a cross-attention fusion layer between two adjacent blocks. Assuming that each block contains m layers, for the first layer of block b (Layer bm), the input prompt is reconstructed from the output of the last layer of block $b-1$ (Layer $bm-1$):

$$S_{vp}^{bm} = \text{softmax}(\frac{(H_{lp}^{bm-1}W_q)(H_{vp}^{bm-1}W_k)^\top}{\sqrt{d}})(H_{vp}^{bm-1}W_v), \tag{19}$$

$$S_{lp}^{bm} = \text{softmax}(\frac{(H_{vp}^{bm-1}W_q)(H_{lp}^{bm-1}W_k)^\top}{\sqrt{d}})(H_{lp}^{bm-1}W_v), \tag{20}$$

where S_{vp}^{bm} and S_{lp}^{bm} are the input for Layer bm.

Intuitively, an efficient prompt fusion should introduce knowledge from another modality while retaining the original specialization. We manage it through controlling the number of blocks in BAF. With an appropriate block number, the representations of different modality get fused gradually as the blocks progress, facilitating a seamless transition between two stages and striking a balance between single-modal specialization and multi-modal fusion.

5. Experimental Setup

5.1. Dataset and Evaluation Metrics

We conduct experiments on two representative MSU tasks: multi-modal sarcasm detection and multi-modal sentiment analysis.

Multi-modal Sarcasm detection For image-text MSD, the MSDT dataset (Cai et al., 2019) stands as the sole benchmark dataset currently available. MSDT has 29k/2.4k/2.4k sample pairs for train/validation/test, each of which contains an image-text pair with a binary label {sarcasm, non-sarcasm}. We keep the test set unchanged and randomly select 32 samples from the train/validation set to construct our few-shot dataset. To balance the label distribution, we control the samples for each label to account for half of the total number. Following the previous work, we adopt Accuracy and F1 as our evaluation metrics. To improve measuring robustness, we sample three disjoint datasets and report the mean result on them. The statistics of MSDT dataset is shown in Table 2.

Multi-modal Sentiment Analysis Our experiments for MSA are based on the MVSA-S (Niu et al., 2016) dataset. Each sample in MVSA-S contains an image-text pair annotated with one of the labels: {positive, neutral, negative}. For few-shot MSA, Yu et al. (2022) performed random sampling on both training and development sets from MVSA-S, constituting 1% of the total. We keep the same setting with them for fair comparison and use Accuracy, Weighted-F1, and Macro-F1 as metrics.

5.2. Implementation Details

We choose VLMo-Base-plus as our baseline model. For data pre-processing, we follow the same steps as VLMo. The images are resized to 224×224 resolution and segmented into 16×16 patches with RandAugment (Cubuk et al., 2020). We utilize the tokenizer from the uncased version of BERT,

	Full Split		Few-shot Split		Avg. length
	#Sarcasm	#Nonsarcasm	#Sarcasm	#Nonsarcasm	
Train	8642	11174	16	16	21.85
Dev	959	1451	16	16	21.79
Test	959	1450	959	1450	22.22

Table 2: Statistics of MSDT dataset, with Avg. length calculated using Bert Tokenizer.

limiting the input text to a length of 40 and truncating any exceeding portions. For hyper-parameters, the block number is set to 2, and the prompt length is set to 10 by default. All experiments adhere to the full parameter training paradigm. During training, the model is optimized by AdamW (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9, \beta_2 = 0.998$. The peak learning rate is $3e-5$. Weight decay is 0.01. In the 32-shot setting, the batch size is 8, and we use linear warmup over the first 10% of the whole 200 steps. We use 1 Nvidia GeForce 3080Ti card for experiments. One training needs around 8GB GPU memory and takes about 2 hours.

5.3. Comparing Methods

We conduct a comprehensive evaluation by comparing with the vanilla multi-modal pre-trained methods, the methods specialized for MSD or MSA tasks, and the prompt methods applied to VLMs.

Multi-modal Pre-trained methods

CLIP (Radford et al., 2021) holds a dual-encoder architecture, designed to encode texts and images separately. We further add trainable projection layers at the output of both encoders and perform classification on their concatenated encoding results. **InstructBLIP** (Dai et al., 2023) is the latest and most advanced general-purpose VLM that introduces instruction tuning techniques to extract informative features tailored to the given instruction. We adopt the Vicuna7B as the base and design three textual prompts for MSD and MSA as the instruction (shown in Table 3) and report their average performance. We use 1 Nvidia A40 card for training InstructBLIP and it takes around 24GB GPU memory.

Multi-modal Prompt Learning Methods

We also consider the multi-modal prompt methods based on CLIP as our baselines. **CoOp** (Zhou et al., 2022b) replaces the manual-crafted text prompts in CLIP with continuous vectors. **MaPLe** (Khattak et al., 2023) leverages prefix tuning in both modality encoders and designs a coupling function to enable prompt interactions. **CMPA** (Liu et al., 2023b) designs prompts for both encoders and employs a cross-modal prompt attention at each layer.

Multi-modal Sarcasm Detection methods

For the MSD task, previous work has only considered a full-shot experimental setup. Based on our few-shot dataset, we replicate some of the studies for which the source code is available. **HFM** (Cai et al., 2019) leverages a hierarchical strategy to fuse the image, attribute and text modalities by attention weights. **ResBERT** (Pan et al., 2020) captures the cross-modal incongruity by inter-modal attention and contradiction within the text by co-attention. **HKE** (Liu et al., 2022a) learns composition-level congruity based on graph neural networks and introduces auto-generated captions as external knowledge.

Multi-modal Sentiment Analysis methods

MVAN (Yang et al., 2021a) utilizes a multi-view memory network to extract single-modal emotion features and interactively capture the cross-view dependencies between the image and text. **MGNNS** (Yang et al., 2021b) learns multi-modal representations by a multi-channel graph neural network based on the global characteristics of MVSA. **UPMPF** (Yu et al., 2022) adopts the multi-modal prompt-based finetuning paradigm and proposes a pre-training stage to narrow the semantic gap between image and text encoders.

Different Prompting Learning Strategies

Manual Prompt defines a hard template combined with input. The representation of [MASK] is then fed into the verbalizer for prediction. **Soft Prompt** prepends several virtual tokens before the input and utilizes [CLS] token for classification. **P-Tuning** (Liu et al., 2021b) combines soft prompt with manual prompt. Table 3 lists the templates used in these methods. **P-Tuning v2** (Liu et al., 2021a) adopts the prefix-tuning idea and expands the prompt parameter space to each transformer layer, using [CLS] token for classification.

6. Results and Analysis

6.1. Evaluation on Multi-modal Sarcasm Detection

We conduct extensive experiments on the dataset MSDT, and report the results in Table 4. We im-

	MSD	MSA
Manual Prompt Soft Prompt P-Tuning	The image-text pair is [MASK]. <text> [V1] [V2] ... [Vn] <text> [V1] [V2] ... [Vn] The image-text pair is [MASK]. <text>	Sentiment of the text: [MASK]. <text> [V1] [V2] ... [Vn] <text> [V1] [V2] ... [Vn] Sentiment of the text: [MASK]. <text>
Instruction Prompt	1.Text:<text> Answer the question: Is this image-text pair sarcastic or nonsarcastic? Answer: 2.Text:<text> Based on the image and text, answer the question: Is this image-text pair sarcastic or nonsarcastic? Answer: 3.Text: <text> Combining the text, is this sarcastic or nonsarcastic? Answer:	1.Text:<text> Answer the question: Which sentiment does this image-text pair contain, negative, neutral or positive? Answer: 2.Text:<text> Based on the image and text, answer the question: Which sentiment does this image-text pair contain, negative, neutral or positive? Answer: 3.Text:<text> Combining the text, which sentiment does this contain, negative, neutral or positive? Answer:

Table 3: The text templates used in different prompt methods for MSD and MSA tasks. The instruction prompt is for InstructBLIP. We do not include P-Tuning v2 here as it does not require a template for input texts.

Methods	w/ CH	w/ LMH	Accuracy	Precision	Recall	F1
HFM (Cai et al., 2019)	✓		58.88 (2.93)	48.93 (2.00)	65.83 (12.28)	55.73 (4.51)
resBERT (Pan et al., 2020)	✓		58.85 (2.04)	48.95 (1.62)	70.87 (4.61)	57.82 (0.45)
HKE (Liu et al., 2022a)	✓		60.76 (4.00)	52.56 (3.41)	74.43 (14.13)	61.02 (3.80)
VLMo (Bao et al., 2022)	✓		60.18 (3.62)	50.97 (4.26)	60.31 (5.28)	53.96 (3.65)
CLIP (Radford et al., 2021)	✓		61.31 (1.28)	51.24 (1.30)	68.33 (8.26)	58.26 (2.22)
InstructBLIP (zero-shot)		✓	57.59 (2.87)	47.22 (3.12)	50.33 (13.25)	47.89 (5.49)
InstructBLIP (Dai et al., 2023)		✓	60.56 (4.02)	50.57 (3.29)	**78.13** (5.94)	61.05 (2.11)
CoOp (Zhou et al., 2022b)	✓		63.00 (5.99)	53.06 (6.42)	67.71 (4.69)	59.41 (5.23)
CMPA (Liu et al., 2023b)	✓		59.94 (2.34)	49.75 (2.35)	63.43 (4.08)	55.75 (2.86)
MaPLE (Khattak et al., 2023)	✓		61.28 (2.43)	50.87 (1.95)	78.00 (6.42)	61.26 (2.80)
VLMo + Manual Prompt		✓	59.85 (2.77)	50.12 (2.64)	73.20 (8.25)	59.10 (1.15)
+ Soft Prompt	✓		62.74 (0.65)	53.26 (1.56)	57.84 (11.09)	54.74 (4.71)
+ P-Tuning (Liu et al., 2021b)		✓	60.34 (1.61)	50.11 (1.31)	75.46 (4.03)	60.21 (2.06)
+ P-Tuning v2 (Liu et al., 2021a)	✓		61.78 (3.02)	52.23 (2.94)	63.99 (15.12)	56.36 (5.98)
VLMo + MoPE-BAF	✓		64.06 (0.71)	53.69 (4.87)	71.60 (2.78)	61.32 (2.57)
+ MoPE-BAF + MP		✓	**65.32** (3.30)	**55.92** (3.91)	68.64 (7.52)	**61.32** (1.15)

Table 4: Experimental results on the MSDT dataset with 32 training samples, where the standard deviations are shown in parentheses. CH refers to the classification head and LMH refers to the language modeling head. MP means Manual Prompt.

plement two settings for our method. The first employs a classification head using [CLS] for prediction, termed MoPE-BAF. The second uses an additional manual template containing [MASK], termed MoPE-BAF + Manual Prompt (our full model).

Compared to previous methods that target full-shot MSD, our model MoPE-BAF achieves significant improvements over HFM and resBERT, and surpasses HKE by a margin of 4.56 points on Accuracy, which leverages external caption knowledge.

Among the current advanced multi-modal models, our method achieves promising results. Compared with the backbone model VLMo, we realize 5.14 and 7.36 points improvement on Accuracy and F1, respectively. Compared with CLIP, our methods attain consistent improvements across all four metrics. In addition, our model with only 150m parameters surpasses the large model InstructBLIP with 8.2B parameters, by 4.76 points increment on Accuracy. This efficiency highlights the superiority of our methods and suggests our potential for practical resource-saving applications.

In comparison to the multi-modal prompt learning methods CoOp, MaPLE and CMPA that implemented on CLIP, MoPE-BAF showcases outstanding performance despite being built on a less powerful backbone model VLMo. Besides, under the scope of prompt methods on VLMo, our full model achieves the best results, whether combined with a classification head or a LM head with the verbalizer. We speculate that our method distinguishes prompts for different modalities and facilitates the interaction between them as the layers deepen, which enhances the specialization of single-modal representation and guarantees a thorough modality fusion. Thus, we obtain significant improvements compared with the non-specific prompt methods P-Tuning and P-Tuning v2, even though P-Tuning v2 requires a larger prompt parameter space for prepending soft prompts in each Transformer layer.

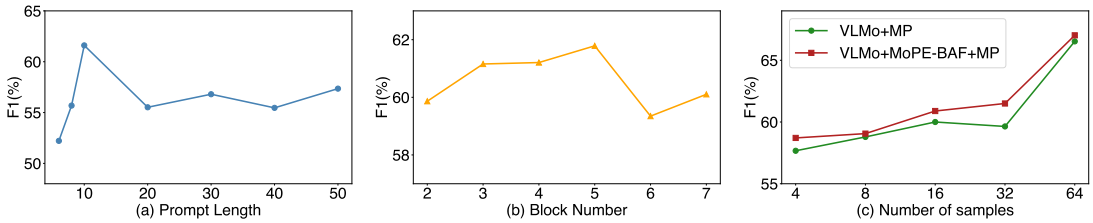

Figure 4: (a) F1 performance training MoPE with different prompt lengths. (b) F1 scores training MoPE-BAF with different block numbers. (c) Comparison between VLMo and VLMo + MoPE-BAF under different training shots.

Methods	Acc	Mac-F1	Wtd-F1
MVAN	42.77	36.75	44.14
MGNNS	34.4	32.05	36.9
UP-MPF	58.21	51.08	58.49
CLIP	49.51	45.67	51.63
CoOp	51.47	40.58	48.52
MaPLe	50.49	43.06	51.74
CMPA	56.74	42.75	53.86
InstructBLIP	59.80	48.59	59.73
VLMo + MP (zero-shot)	59.07	38.53	51.94
VLMo + MP	60.79	52.62	61.27
VLMo + P-Tuning	61.03	51.28	60.75
VLMo + MoPE-BAF + MP	**63.48**	**52.92**	**62.40**

Table 5: Comparison of results between our approach and previous methods on the few-shot MVSA-S dataset. Mac-F1 and Wtd-F1 denote Macro-F1 and Weighted-F1 respectively. MP denotes manual prompt. The results of the first group are from the work (Yu et al., 2022), and we implemented the models in the second group.

Methods	A	P	R	F1
VLMo	60.18	50.97	60.31	53.96
+ MoPE	61.73	51.35	**75.13**	60.94
+ MoPE + BAF	**64.06**	**53.69**	71.60	**61.32**
VLMo + MP	59.85	50.12	**73.20**	59.10
+ MoPE	64.06	53.98	70.73	60.93
+ MoPE + BAF	**65.32**	**55.92**	68.64	**61.32**

Table 6: Ablative analysis of the MoPE and BAF modules.

6.2. Evaluation on Multi-modal Sentiment Analysis

For the MSA task, the experimental results on MVSA-S dataset are shown in Table 5. There is a huge performance gap between the non-pretrained methods (MVAN, MGNNS) and pre-trained methods. Among the pre-trained multi-modal methods, InstructBLIP demonstrates impressive efficacy, outperforming the current state-of-the-art UP-MPF. VLMo series perform better than CLIP series in this task, perhaps because the unified structure of VLMo, which effectively leverages both text and image input, is well-suited to this type of task, and the abundant pre-training tasks (especially mask language modeling), endows it with the ability to infer sentiment.

VLMo performs well in the zero-shot setting, establishing a robust baseline. When integrated with prompts, the performance of VLMs is significantly boosted. VLMo adopting prompt-based finetuning (+MP) delivers the most superior results compared to those without VLMo, even outperforming the large-scale language model InstructBLIP. After combining with our MoPE-BAF method, the performance is further elevated, surpassing UP-MPF significantly by 3.91 F1 points. It demonstrates the generality and validity of our proposed model in different MSU tasks.

6.3. Ablation Study

We conduct ablation experiments to investigate the effects of MoPE and BAF, and the results are presented in Table 6. In both settings (with or without using Manual Prompt), MoPE significantly improves the performance of VLMo, and further adding the prompt fusion operation achieves better performance. Please note that we cannot validate the effect of BAF separately since it cannot exist independently of MoPE.

6.4. Model Analysis

We also conduct experiments to analyze some controllable factors in the proposed MoPE-BAF model.

Prompt Length Generally, a longer prompt correlates with an increase in learnable parameters, while in the few-shot setting, it may exacerbate the over-fitting problem. We vary the prompt length from 5 to 50 in MoPE, and the results concerning different lengths are visualized in Figure 4 (a). Overall, the impact of prompt length on model performance shows a trend of initially increasing, subsequently diminishing, and then stabilizing. Our explanation is that prompts too short are insufficient to bring about qualitative changes to the model, while too long may pose challenges for searching for the opti-

mal solution due to the complexity of the expanded search space.

Block Number We investigate the impact of the block number in the BAF module. Specifically, in the first 21 layers of VLMo that use V-Prompt and L-Prompt experts, we divide them into 2-7 blocks. A configuration with 1 block implies no prompt fusion, and is therefore not displayed here. If the number of model layers cannot be evenly divided by the block number, the excess layers are allocated to the bottom blocks. This ensures that the maximum layer difference between any two blocks does not exceed 1. The results are shown in Figure 4 (b). When the block number is between 3 and 5, the overall performance remains almost similar. However, a noticeable decline can be observed after 6 blocks. We speculate that too many cross-modal cross-attention operations between blocks cause their original specialization to be discarded, thereby contradicting the function of MoPE. An appropriate number of blocks can strike a balance between the specialization and different modalities interaction.

Training Shots The number of training shots plays a crucial role in model performance, particularly in the few-shot setting. We conduct experiments to train the VLMo base and our full model with different numbers of samples. The results are presented in Figure 4 (c). We find that the performance of both models improves as training shots increase. Besides, applying MoPE-BAF consistently improves the performance of VLMo with the number of training examples from 4 to 64 in the f1 score, which demonstrates the robustness and efficiency of our method.

7. Conclusion

In this paper, we present MoPE-BAF, a new multi-modal soft prompt framework catering to unified VLMs for few-shot multi-modal tasks. Specifically, we devise two prompt experts to serve the text and image modality separately with a better specialization ability, and further activate the interactions of prompt experts by inserting cross-modal prompt attention between adjacent Transformer blocks. In this way, we reach a harmonious balance in modality specialization and fusion, thus fulfilling the requirement of modeling deep relations between modalities. Experiments on multi-modal sarcasm detection and multi-modal sentiment analysis show that our MoPE-BAF model not only surpasses other widely-used prompt methods, but also outperforms the advanced large language models. Further analysis confirms the effectiveness of MoPE and BAF.

In the future, we intend to incorporate task-related external knowledge into our prompt design, and broaden the scope of our method to include other tasks, such as multi-modal content generation and multi-modal reasoning.

8. Limitations

While our study provides insights into the soft prompt technique on VLMs, it has to be acknowledged that it has some limitations. First, we observe a sensitivity of MoPE during training. The performance of MoPE relies on appropriate hyperparameter selection and the optimal hyperparameters differ across downstream tasks. While performing perform a grid search on hyperparameters may mitigate the issue, we believe that how to effectively control the sensitivity is worth further exploration in future work.

Besides, although MoPE-BAF can theoretically be applied to any unified VLMs without disrupting the architecture or encoding process of the base architecture, our study was conducted exclusively on VLMo, and we have not yet extended MoPE-BAF to other pre-trained VLMs. This restricts the generalizability analysis of our methods. In the future, we would apply MoPE-BAF on more VLMs, which we think would provide a more comprehensive understanding of the role and impact of our methods.

Acknowledgement

This work is supported by the Key Project of Natural Science Foundation of China (61936012) and the National Natural Science Foundation of China (62076008).

Bibliographical References

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, volume 35, pages 32897–32912. Curran Associates, Inc.

Alexandru-Costin Băroiu and Ștefan Trăușan-Matu. 2022. Automatic sarcasm detection: Systematic literature review. *Information*, 13(8).

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10563–10571.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021a. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 6–15, New York, NY, USA. Association for Computing Machinery.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021b. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*.

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

Tony Huang, Jack Chu, and Fangyun Wei. 2022. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of*

Machine Learning Research, pages 5583–5594. PMLR.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4707–4715, New York, NY, USA. Association for Computing Machinery.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.

Hui Liu, Wenya Wang, and Haoliang Li. 2022a. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too.

Xuejing Liu, Wei Tang, Jinghui Lu, Rui Zhao, Zhaojun Guo, and Fei Tan. 2023b. Deeply coupled cross-modal prompt learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7957–7970, Toronto, Canada. Association for Computational Linguistics.

Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. 2022. Vision-and-language pretrained models: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5530–5537. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rishabh Misra and Prahal Arora. 2023. Sarcasm detection using news headlines dataset. *AI Open*, 4:13–18.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multiview social data. In *MultiMedia Modeling*, pages 15–27, Cham. Springer International Publishing.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online. Association for Computational Linguistics.

Shraman Pramanick, Aniket Roy, and Vishal M. Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3930–3940.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Suyash Sangwan, Md Shad Akhtar, Pranati Behera, and Asif Ekbal. 2020. I didn't mean what i wrote! exploring multimodality for sarcasm detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and LiangLiang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 1136–1145, New York, NY, USA. Association for Computing Machinery.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2468–2480, Toronto, Canada. Association for Computational Linguistics.

Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. 2022. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8164–8175, Dublin, Ireland. Association for Computational Linguistics.

Colin Wei, Sang Michael Xie, and Tengyu Ma. 2021. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. In *Advances in Neural Information Processing Systems*, volume 34, pages 16158–16170. Curran Associates, Inc.

Changsong Wen, Guoli Jia, and Jufeng Yang. 2023. Dip: Dual incongruity perceiving network for sarcasm detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2540–2550.

Yang Wu, Yanyan Zhao, Xin Lu, Bing Qin, Yin Wu, Jian Sheng, and Jinlong Li. 2021. Modeling incongruity between modalities for multimodal sarcasm detection. *IEEE MultiMedia*, 28(2):86–95.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 929–932, New York, NY, USA. Association for Computing Machinery.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786, Online. Association for Computational Linguistics.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2021a. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021b. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339, Online. Association for Computational Linguistics.

Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 189–198, New York, NY, USA. Association for Computing Machinery.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. 2022. Unified vision and language prompt learning.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023a. Vision-language models for vision tasks: A survey.

Yazhou Zhang, Dan Ma, Prayag Tiwari, Chen Zhang, Mehedi Masud, Mohammad Shorfuzzaman, and Dawei Song. 2023b. Stance-level sarcasm detection with bert and stance-centered graph attention networks. *ACM Trans. Internet Technol.*, 23(2).

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *Int. J. Comput. Vision*, 130(9):2337–2348.

Language Resource References

Cai, Yitao and Cai, Huiyu and Wan, Xiaojun. 2019. *Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model*. Institute of Computer Science and Technology, Peking University.

Teng Niu and Shiai Zhu and Lei Pang and Abdulmotaleb El-Saddik. 2016. *MVSA: Sentiment Analysis on Multi-view Social Data*. MCRLab, University of Ottawa.