Continuous Attentive Multimodal Prompt Tuning for Few-Shot Multimodal Sarcasm Detection

Soumyadeep Jana, **Animesh Dey**, and **Sanasam Ranbir Singh**
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
{sjana, d.animesh ,ranbir}@iitg.ac.in

Abstract

With the steep rise in multimodal content on social media, multimodal sarcasm detection has gained widespread attention from research communities. Existing studies depend on large-scale data, which is challenging to obtain and expensive to annotate. Thus, investigating this problem in a few-shot scenario is required. Overtly complex multimodal models are prone to overfitting on in-domain data, which hampers their performance on out-of-distribution (OOD) data. To address these issues, we propose **C**ontinuous **A**ttentive **M**ultimodal **P**rompt Tuning model (CAMP), that leverages the prompt tuning paradigm to handle few-shot multimodal sarcasm detection. To overcome the siloed learning process of continuous prompt tokens, we design a novel, continuous multimodal attentive prompt where the continuous tokens intricately engage with both image and text tokens, enabling the assimilation of knowledge from different input modalities. Experimental results indicate that our method outperforms other multimodal baseline methods in the few-shot setting and OOD scenarios. Our few-shot dataset and code is available at `https://github.com/mr-perplexed/camp`.

1 Introduction

Sarcasm is a figurative language where the utterance conveys a meaning opposite to the literal meaning of the words used. Detecting sarcasm is important for effectively understanding sentiment (Maynard and Greenwood, 2014; Badlani et al., 2019), hate speech (Frenda, 2018; Yang et al., 2022a), and users' opinions on social media (Tindale and Gough, 1987; van Eemeren and Grootendorst, 1992; Averbeck, 2013; Ghosh et al., 2021). With the rise in multimodal content on social media platforms, multimodal sarcasm detection has gained widespread attention from research communities. Multiple modalities provide a crucial clue to ascertain the sarcastic nature of a post since deciphering sarcasm from uni-modal (only text or image) content may be highly ambiguous or unspecified.

Current approaches for multimodal image-text sarcasm detection (Cai et al., 2019; Pan et al., 2020; Xu et al., 2020; Liang et al., 2021; Liu et al., 2022a; Liang et al., 2022; Tian et al., 2023; Wen et al., 2023) suffer from some major challenges. These models primarily rely on large annotated datasets to achieve good performance. However, these datasets are difficult to obtain, and annotation is expensive and highly challenging due to socio-cultural and contextual dependencies (Rockwell and Theriot, 2001; Ivanko and Pexman, 2003; Dress et al., 2008; Oprea and Magdy, 2019). Distant supervision techniques for labeling, like the use of special markers such as #sarcasm on Twitter, introduce additional noise in the form of wrong labels (Davidov et al., 2010; González-Ibáñez et al., 2011). Due to the complex structure of these multimodal models, they tend to overfit on in-domain data causing a reduction in performance for out-of-distribution (OOD) data.

Prompt-based methods have gained popularity in few-shot learning as they enable Pretrained Language Models (PLMs) to generalize to new tasks with minimal or no training data as PLMs can serve as knowledge bases (Petroni et al., 2019; Jiang et al., 2020) due to their large-scale training on huge corpora. Hence, it is imperative to use prompt-based method for our task.

Most of the existing prompt-based works on downstream tasks are based on prompt-based finetuning (Cao et al., 2022; Yang et al., 2022a,b; Yu and Zhang, 2022), where discrete prompts are given as inputs to PLMs, and the entire PLM is finetuned to fill up the mask token. This poses three main challenges. First, finding the right prompt in the discrete token space is difficult and often yields sub-optimal performance. Changes in token count drastically impact results (Liu et al., 2021).

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 314–326
November 15-16, 2024 ©2024 Association for Computational Linguistics

Second, training all model weights increases parameters, memory use, and training time. Lastly, fine-tuning pre-trained language models often leads to catastrophic forgetting (Wang et al., 2022; Zhai et al., 2023), reducing generalizability and performance on out-of-distribution data due to changes in the pre-trained weights.

Motivated by the shortcomings of traditional multimodal approaches and discrete prompt-based techniques, we explore the idea of Prompt Tuning (Li and Liang, 2021; Lester et al., 2021), a new paradigm involving PLMs where task-specific continuous prompts are learned during training, keeping the parameters of the PLM frozen. Further, in the vanilla Prompt Tuning approach, a significant limitation arises from the frozen nature of the pre-trained language model (PLM). This constraint results in independent learning of continuous prompt tokens without integrating knowledge on how to attend to both image and text tokens effectively. To address this challenge, we propose a novel model: **CAMP** (**C**ontinuous **A**ttentive **M**ultimodal **P**rompt Tuning) model for few-shot multimodal sarcasm detection. To begin with, we design multimodal continuous prompts with text and image modalities. We also use captions of the images as the third modality to bridge the semantic gap between image and text. Our approach enhances the model's ability to better learn the continuous prompt tokens by incorporating multimodal information and introducing attentive mechanisms, thereby significantly improving its capacity to attend to both image and text tokens seamlessly.

Our results show that using only 0.3 fraction of the entire PLM parameters, CAMP can achieve state-of-the-art results in few-shot multimodal sarcasm detection. CAMP also shows strong performance on the OOD setting. In summary, the main contributions and findings of this paper are listed below:

1. To the best of our knowledge, this study is the first to investigate multimodal sarcasm detection in a few-shot setup using continuous prompt tuning paradigm.

2. We propose **CAMP**, a parameter efficient model leveraging novel continuous attentive multimodal prompt.

3. Our extensive experiments on two benchmark datasets showcase our model's superiority over strong multimodal baselines in a few-shot and OOD setting.

4. We present a comprehensive analysis of different prompt-based techniques including prompting, prompt-based finetuning, and prompt tuning on our task.

2 Related Work

2.1 Multimodal Image-Text Sarcasm Detection

The field of sarcasm detection started with text as the sole modality. Prior works (Joshi et al., 2015; Khattri et al., 2015; Joshi et al., 2016; Amir et al., 2016; Zhang et al., 2016; Poria et al., 2016; Ghosh et al., 2017; Agrawal and An, 2018; Agrawal et al., 2020; Babanejad et al., 2020; Lou et al., 2021; Liu et al., 2022b) use different sequence modeling techniques, along with external cues like author information, conversation context, etc, to detect the incongruity present in the text. With the rise in the usage of multimodal content on social media, researchers shifted their attention towards multimodal sarcasm detection. (Schifanella et al., 2016) was the first to perform the task of multimodal sarcasm detection with text and image modality. This work used manually designed features to detect incongruity between the two modalities. (Cai et al., 2019) released a new image-text dataset based on Twitter and proposed a hierarchical early and late fusion method to combine the two modalities. Work by (Xu et al., 2020) employed decomposition and relation network to identify cross-modality incongruity and semantic association. Study by (Pan et al., 2020) showed that sarcasm could arise from either intra-modal or inter-modal associations. So, they proposed a self-attention-based model to capture intra and inter-modal incongruity. (Liang et al., 2021, 2022) used graph neural networks over in-modal and cross-modal graphs to detect sarcasm. To model both granular-level and abstract-level incongruities, (Liu et al., 2022a) used hierarchical semantic interactions between image-text modalities. (Wen et al., 2023) proposed a Dual Incongruity Perceiving (DIP) network, which combines semantic intensified distribution modeling and siamese sentiment contrastive learning modules to distinguish between sarcastic and non-sarcastic samples. (Tian et al., 2023) proposed a Dynamic Routing Transformer model to adaptively capture the inter-modal contrast between image and text to identify sarcasm.

Unlike traditional methods relying on extensive annotated data and training of PLMs like BERT (Devlin et al., 2019) as foundational components, our approach operates in a few-shot learning scenario, utilizing a frozen PLM. This strategy proves effective in handling the scarcity of sarcasm annotations while achieving state-of-the-art performance with only a fraction of the PLM parameters.

2.2 Multimodal Prompt-Based Approaches

Recent studies have used prompt-based methods for various multimodal NLP downstream tasks like visual QA (Liu et al., 2022c; Chappuis et al., 2022; Guo et al., 2022; Ossowski and Hu, 2023), sentiment analysis (Gao et al., 2021; Yang et al., 2022b; Yu et al., 2022; Yu and Zhang, 2022; Hosseini-Asl et al., 2022), and hate speech detection (Cao et al., 2022; Ji et al., 2023; García-Díaz et al., 2023; Cao et al., 2023).

Most of these approaches have either focused on prompting or prompt-based finetuning paradigms. However, a detailed study on using continuous prompts for multimodal sarcasm detection is yet to be explored. To this end, we propose a continuous prompt tuning approach to tackle multimodal sarcasm detection with attentive prompts.

3 Proposed Approach

3.1 Problem Definition

Given a multimodal sample $x_j = (T_j, I_j)$, where $T_j = \{t_j^1, t_j^2, ..., t_j^n\}$ is the text and I is the associated image, the task is to assign x_j a label $y_j \in Y = \{sarcastic, nonsarcastic\}$. Traditionally, the task of multimodal sarcasm detection has been formulated as a binary classification task, wherein the model outputs two probabilities corresponding to the label space $Y = \{sarcastic, nonsarcastic\}$. The sample is classified based on the higher probability label. We reformulate the task as a Masked Language Modeling Problem. Given a PLM M, M is prompted with multimodal input to fill the $[MASK]$ token, which represents the labels Y.

3.2 Multimodal Prompt Tuning

We propose a novel model called **CAMP** (**C**ontinuous **A**ttentive **M**ultimodal **P**rompt Tuning) model for few-shot multimodal sarcasm detection. Figure 1 shows the overall architecture of our proposed model. In this sub-section, we elaborate on the design of continuous multimodal prompt, while in the next sub-section, we delve into incorporating attention mechanism into the continuous prompt tokens to generate continuous attentive multimodal prompt.

Given a multimodal sample consisting of text T_j and an associated image I_j, text modality T_j can be directly fed to the PLM. However, PLMs are not designed to accommodate image modality information. To curb this, following (Yang et al., 2022b), we generate pseudo-visual tokens. First, the original image I_j is passed through ResNet, and it is then projected into the text feature space using a weight matrix W^t and bias vector b^t, as depicted by the equation:

$$V_j = W^t * ResNet(I_j) + b^t \tag{1}$$

The V_j is then reshaped, $V_j = reshape(V_j) = \{v_j^1, v_j^2, ..., v_j^p\}$, where $V_j \in \mathbb{R}^{p \times v_{dim}}$ to generate the final visual tokens where p is the number of image token slots and is kept as a hyperparameter. After introducing the visual tokens, to further reduce the gap between image and text modalities, we generate caption C_j using a vision-language model BLIP-2 (Li et al., 2023), where $C_j = BLIP2(I_j)$. BLIP-2 combines frozen pre-trained image models with language models for representation and generative learning. This helps BLIP to achieve state-of-the-art performance in image captioning task. With these at our disposal, we design our multimodal prompt template Z as follows which can be fed to the PLM for it to generate the $[MASK]$ token:

$$\boxed{\begin{array}{l} Z(T_j, C_j, V_j) = [V_j]\ Tweet\ text : [T_j] \\ Caption : [C_j].\ [MASK] \end{array}}$$

Subsequently, the PLM embeds Z as a series of m discrete tokens by passing through its encoder, creating an embedding matrix $F \in \mathbb{R}^{m \times h_{dim}}$.

Now, we design our continuous prompts. In the prompt tuning paradigm introduced in (Li and Liang, 2021), learnable vectors called continuous prompt tokens are added to the prompt being fed to PLM. These continuous tokens are generated from a prompt encoder, particularly multilayer perceptron or LSTM networks. During training, instead of fine-tuning the PLM, these continuous tokens are learned for the task at hand. This differs from the approach of prompting or prompt-based finetuning. In prompting, discrete prompt tokens are employed to query the PLM without modifying the PLM, while in prompt-based finetuning, all the PLM weights are updated. Figure 2 presents a schematic difference between the paradigms.

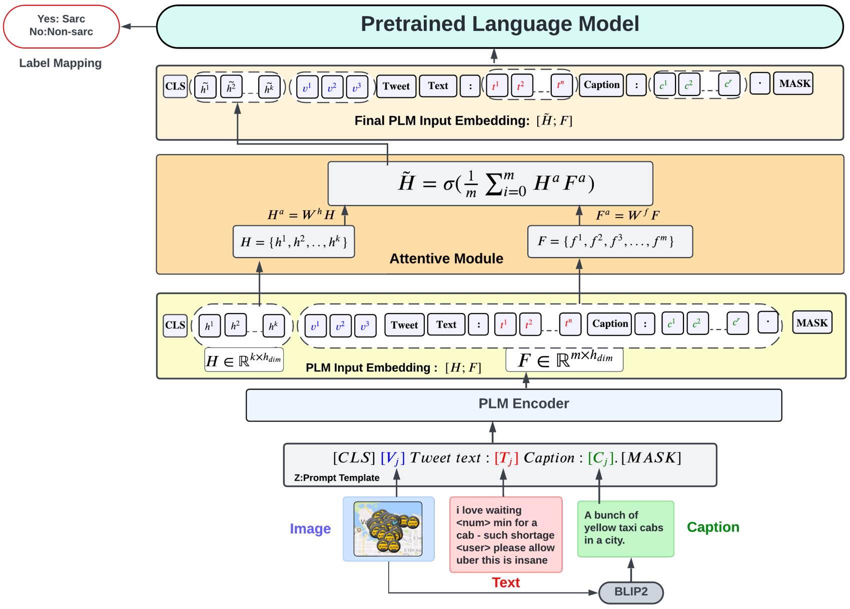
Yes: Sarc
No:Non-sarc
Label Mapping
Pretrained Language Model
Final PLM Input Embedding: $[\tilde{H}; F]$
$\tilde{H} = \sigma(\frac{1}{m}\sum_{i=0}^{m} H^a F^a)$
$H^a = W^h H$
$F^a = W^f F$
$H = \{h^1, h^2, .., h^k\}$
$F = \{f^1, f^2, f^3, ..., f^m\}$
Attentive Module
$H \in \mathbb{R}^{k \times h_{dim}}$
PLM Input Embedding : $[H; F]$
$F \in \mathbb{R}^{m \times h_{dim}}$
CLS
Tweet
Text
Caption
MASK
PLM Encoder
$[CLS]\ [V_j]\ Tweet\ text : [T_j]\ Caption : [C_j].[MASK]$
Z:Prompt Template
Image
i love waiting <num> min for a cab - such shortage <user> please allow uber this is insane
Text
A bunch of yellow taxi cabs in a city.
Caption
BLIP2

Figure 1: Architecture of our CAMP model.

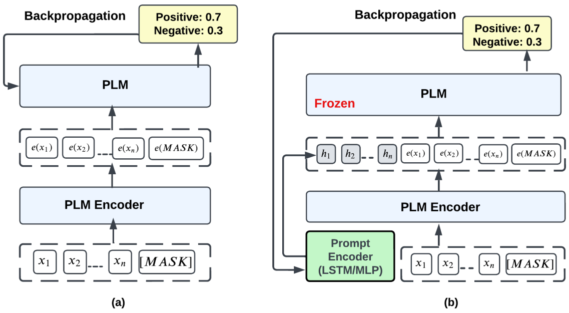
Backpropagation
Positive: 0.7
Negative: 0.3
PLM
$e(x_1)$ $e(x_2)$ $e(x_n)$ $e(MASK)$
PLM Encoder
x_1 x_2 x_n $[MASK]$
(a)
Backpropagation
Positive: 0.7
Negative: 0.3
Frozen
PLM
h_1 h_2 h_n $e(x_1)$ $e(x_2)$ $e(x_n)$ $e(MASK)$
PLM Encoder
Prompt Encoder (LSTM/MLP)
x_1 x_2 x_n $[MASK]$
(b)

Figure 2: Schematic Representation of a) Prompt-Based Finetuning strategy and b) Prompt Tuning strategy.

We prepend the continuous learnable tokens to the prompt, which are represented by the matrix $H = \{h_1, h_2, ..., h_k\} \in \mathbb{R}^{k \times h_{dim}}$, where k is the number of continuous prompt tokens. The parameters of the underlying prompt encoder is represented by ϕ. H is then combined with the embedded input F, resulting in a unified matrix $[H; F]$ of dimensions $\mathbb{R}^{(k+m) \times h_{dim}}$. This combined matrix called the PLM input embedding matrix, forms the input for the PLM.

3.3 Attentive Multimodal Prompt Tuning

A significant drawback of the vanilla continuous prompt tuning approach is the siloed learning process for continuous prompt tokens, overlooking the essential integration of knowledge required for effectively attending to both image and text tokens. This happens because the weights of the PLM are frozen in prompt tuning, hindering the function of the attention mechanism. Thus the model cannot focus on specific parts of the input sequence when generating outputs, failing to capture dependencies and relationships between different tokens.

To address this issue, we design a continuous attentive multimodal prompt, where the learnable vectors can attend to the non-learnable or fixed tokens before passing through the PLM layers. We reason that this would capture the dependencies between the learnable and fixed tokens, and act as a substitute for the frozen attention layers of the PLM. We segregate the PLM input embedding matrix $[H; F]$ into two parts, learnable tokens H and non-learnable or fixed token embeddings F, where $H = \{h^1, h^2, .., h^k\}$ and $F = \{f^1, f^2, .., f^m\}$. $[CLS]$ and $[MASK]$ token embeddings are ignored. To find out which learnable tokens attend to which fixed tokens, we parameterize the token embeddings and find out their dot product using the following equations.

$$H^a = W^h H \tag{2}$$

$$F^a = W^f F \tag{3}$$

$$S = H^a F^a, \qquad S \in \mathbb{R}^{k \times m} \tag{4}$$

S denotes the attention scores of the learnable tokens with each of the other fixed tokens.

For learnable token h^l, we calculate its relative attention score $attn^l$ from S.

$$attn^l = \sigma(\frac{1}{m}\sum_{i=0}^{m} S_{li}) \quad (5)$$

We define the new set of learnable tokens as $\tilde{H} = \{\tilde{h}^1, \tilde{h}^2, ..., \tilde{h}^k\}$, where,

$$\tilde{h}^l = attn^l h^l \quad (6)$$

The attentive learnable token matrix $\tilde{H}$ is then combined with the embedded input F, resulting in a unified matrix $[\tilde{H}; F]$ of dimensions $\mathbb{R}^{(k+m)\times h_{dim}}$. This combined matrix forms the final input and is passed through the PLM to generate the $[MASK]$ token.

3.4 Model Training and Prediction

We feed our final input embedding matrix $E = [\tilde{H}; F]$ to the PLM M. The $[MASK]$ token in E helps to recast the problem into a cloze-filling task. The objective of M is to model the probability of predicting class $y_j \in Y$ as:

$$P(([MASK] = y_j)|E) = \frac{e^{W_{y_j} O_{[MASK]}}}{\sum_{y_j \in Y} e^{W_{y_j} O_{[MASK]}}} \quad (7)$$

where $O_{[MASK]}$ is the hidden representation of $[MASK]$ token and W_{y_j} is the final layer weight of the PLM M. The parameters are optimized by using cross-entropy loss. We update the parameters of the continuous vector tokens ϕ, the projection weights, W^h, and W^f during the training process, while the entire set of weights for M is frozen.

4 Experiments

4.1 Datasets

We evaluate our model CAMP on two benchmark datasets MMSD (Cai et al., 2019) and MMSD2.0 (Qin et al., 2023). MMSD2.0 builds upon MMSD by removing spurious cues and re-annotating the unreasonable negative samples. Following (Yu and Zhang, 2022), we randomly sample 1% of the training data with two different seeds for our few-shot setting, keeping the number of samples equal for each category. We maintain $|valid| = |train|$, while the number of samples in the test set is kept the same. The statistics of the dataset are presented in Table 1.

4.2 Experimental Settings

We use BERT-base-uncased as our PLM and NF-ResNet-50 (Brock et al., 2021) as our visual encoder. Both these backbone networks are kept frozen while training. We map the label space of both MMSD and MMSD2.0 datasets from $\{0, 1\}$ to $\{No, Yes\}$, where the label Yes denotes a sarcastic sample. Following (Yu and Zhang, 2022), to account for variation in performance, we experiment three times for each split, totaling 6 (3×2) training runs for each dataset. We report the mean Accuracy (Acc), mean Macro-F1 (F1), and the standard deviation across the 6 runs. We set the batch size to 16 and the learning rate to 1e-4 for both datasets. The number of continuous prompt tokens is set to 50 for MMSD and 80 for MMSD2.0, while image token slots are fixed at 3 for both datasets. The maximum token length for the PLM is 128. We run our model for 20-100 epochs and pick the model that performs best for the validation set for testing. Additional hyperparameter details are in the Appendix section A.1

4.3 Baselines

We compare our proposed model CAMP with four groups of baselines in a few-shot setting.

1. **Text Modality**: We compare with **TextCNN** (Kim, 2014), a CNN based text classification model, and **BiLSTM** (Graves and Schmidhuber, 2005). We finetune standard **BERT** (Devlin et al., 2019) to compare with our model as it uses a BERT-based adaptation. **LM-BFF** (Gao et al., 2021) uses generated text prompts tailored to each dataset and text demonstrations to address few-shot text classification tasks. **LM-SC** (Jian et al., 2022) builds on LM-BFF by incorporating supervised contrastive learning for few-shot text tasks. We also compare a variant of our model **CAMP(w/o img)** without the image and caption tokens.

2. **Image Modality**: Similar to (Cai et al., 2019), we use the image embedding of the pooling layer of **ResNet** (He et al., 2015) for sarcasm classification. We also benchmark on **ViT** (Dosovitskiy et al., 2020), a transformer-based vision model. We also compare a variant of our model **CAMP(w/o txt)** without the text and caption tokens.

3. **Image + Text Modality (Full-Shot)**: We compare our model with state-of-the-art multimodal models for sarcasm detection designed for full dataset setting. **HFM** (Cai et al., 2019) used hierarchical early and late fusion to fuse

	Train			Valid			Test		
Dataset	**Pos**	**Neg**	**Total**	**Pos**	**Neg**	**Total**	**Pos**	**Neg**	**Total**
MMSD	**99** / 8642	**99** / 11174	**198** / 19816	**99** / 959	**99** / 1459	**198** / 2410	**959** / 959	**1450** / 1450	**2409** / 2409
MMSD2.0	**99** / 9572	**99** / 10240	**198** / 19816	**99**/1042	**99** / 1368	**198** / 2410	**1037** / 1037	**1072** / 1072	**2409** / 2409

Table 1: Statistics of MMSD and MMSD2.0 dataset in the few shot setting. For splits presented as **X**/Y, **X** represents the few-shot data sampled while Y represents the total data. The total train split represents approximately 1% of the total training data with $|valid| = |train|$, while the number of samples in the test set is kept the same.

		MMSD		MMSD2.0	
Modality	**Method**	**Acc**	**F1**	**Acc**	**F1**
Image	ResNet	0.664 (0.1)	0.602 (1.2)	0.638 (1.3)	0.625 (0.5)
	ViT	0.611 (1.6)	0.522 (1.7)	0.560 (2.8)	0.614 (0.5)
	CAMP(w/o txt)	0.664 (2.7)	0.635 (3.2)	0.659 (1.9)	0.645 (2.2)
Text	TextCNN	0.631 (2.8)	0.549 (2.5)	0.568 (0.7)	0.570 (1.6)
	BiLSTM	0.602 (1.7)	0.560 (2.3)	0.499 (2.1)	0.595 (2.1)
	BERT	0.667 (2.2)	0.665 (3.1)	0.590 (2.9)	0.623 (2.4)
	LM-BFF	0.695 (2.7)	0.688 (2.3)	0.637 (1.4)	0.626 (2.5)
	LM-SC	0.698 (1.4)	0.681 (0.8)	0.640 (0.7)	0.632 (1.5)
	CAMP(w/o img)	0.696 (1.7)	0.678 (1.5)	0.613 (0.3)	0.560 (2.0)
Image+Text (Full-Shot)	HFM	0.612 (1.3)	0.598 (1.1)	0.561 (0.2)	0.361 (0.3)
	Attn-BERT	0.707 (1.7)	0.696 (1.3)	0.659 (1.6)	0.683 (1.8)
	HKE	0.503 (2.3)	0.667 (2.8)	0.408 (1.5)	0.579 (1.3)
	DIP	0.704 (2.7)	0.698 (2.3)	0.685 (2.8)	0.658 (2.6)
	DynRT	0.583 (0.1)	0.487 (0.6)	0.518 (2.9)	0.513 (3.2)
Image+Text (Few-Shot)	PVLM	0.712 (0.6)	0.699 (0.2)	0.665 (2.2)	0.658 (2.1)
	UP-MPF	0.707 (2.4)	0.701 (2.6)	0.669 (0.4)	0.663 (0.1)
	CAMP(w/o attn)	0.716 (0.5)	0.697 (0.7)	0.662 (0.2)	0.652 (0.4)
	CAMP	**0.729** (0.9)	**0.717** (1.0)	**0.692** (2.8)	**0.681** (2.3)

Table 2: Performance comparison of existing methods with our proposed model CAMP. The best results across metrics are highlighted in bold. Numbers in bracket indicate standard deviation.

image, text, and image attributes. **D&R Net** (Xu et al., 2020) uses semantic association. **Attn-BERT** (Pan et al., 2020) used a self-attention mechanism to model intra and inter-modal incongruity. **InCrossMGs** (Liang et al., 2021) used GCN to model self and cross-modal interaction. A cross-modal image-text GCN is used by **CMGCN**. (Liang et al., 2022) **HKE** (Liu et al., 2022a) used a hierarchical interaction network to model both granular and abstract level incongruities. **DIP** (Wen et al., 2023) network integrates sentiment contrastive learning with semantic modeling. **DynRT** (Tian et al., 2023) used a Dynamic Routing Transformer model.

4. **Image + Text Modality (Few-Shot)**: Due to the lack of few-shot multimodal baselines for our task, we adopt two state-of-the-art baselines from the Multimodal Sentiment Analysis task. **PVLM** (Yu and Zhang, 2022) directly introduces the image features to pre-trained language. **UP-MPF** (Yu et al., 2022) uses pre-training data with tasks based on PVLM. We also compare a variant of our model named as **CAMP(w/o attn)** without the attention module.

We run all the baseline models in their original settings on our few-shot data splits and report the results. The original codes for some of the baselines are not available, and hence we don't include them in our comparisons.[1]

4.4 Main Results

Following (Yu and Zhang, 2022), we report the results on the randomly sampled 1% of the training data in Table 2. Our findings are as follows: (1) CAMP outperforms all other baseline methods for both datasets in unimodal as well as multimodal settings. This demonstrates the efficacy of continuous attentive prompts to leverage pretrained knowledge to classify instances accurately. It can be observed that the performance of CAMP, along with all other baselines, decreases for the MMSD2.0 dataset. This is because certain cues important for sarcasm, like hashtags and emojis, have been completely removed from the text in MMSD2.0. (2) For the unimodal methods, text modality methods perform better than image modality methods in MMSD. This shows that textual features provide more sarcastic cues. (3) For the image modality

[1] Original codes for D&R Net and InCrossMGs are not publicly available, while CMGCN uses extra attributes which is not available.

		MCMD	
Strategy	Method	Acc	F1
Multimodal Baselines	Attn-BERT	0.477 (0.3)	0.474 (0.1)
	DIP	0.545 (1.2)	0.545 (0.8)
	DynRT	0.519 (1.6)	0.518 (1.4)
	PVLM	0.564 (1.8)	0.541 (1.3)
	UP-MPF	0.582 (2.1)	0.577 (1.9)
Prompt-Based Finetuning	PT_1^d	0.578 (0.5)	0.509 (0.8)
	PT_2^d	0.584 (1.7)	0.374 (1.9)
Prompt-Tuning	CAMP(w/o attn)	0.588 (0.4)	0.516 (0.7)
	CAMP	**0.601** (1.3)	**0.591** (1.6)

Table 3: Performance comparison on OOD setting. Discrete Templates PT_i^d used in Prompt-Based Finetuning are listed in Table 6.

	MMSD		MMSD2.0	
Method	Acc	F1	Acc	F1
w/o cap	0.694 (1.3)	0.671 (0.2)	0.655 (1.2)	0.636 (1.7)
w cap	**0.729** (0.9)	**0.717** (1.0)	**0.692** (2.8)	**0.681** (2.3)

Table 4: Ablation on caption tokens for CAMP model

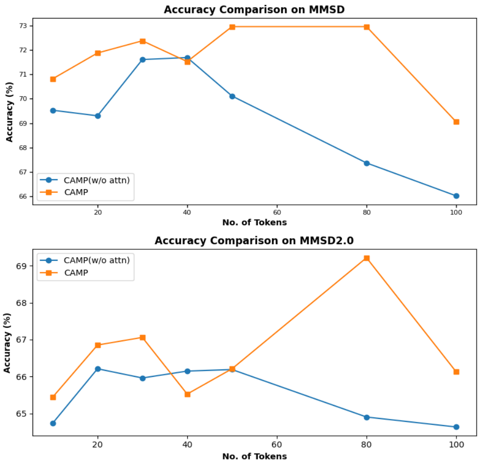

Figure 3: Performance comparison of CAMP and CAMP(w/o attn) over MMSD and MMSD2.0 datasets for various token lengths.

methods, CAMP(w/o text) outperforms other baselines across both the datasets. This observation is interesting because although PLMs are pretrained on text, our attentive, continuous prompt can still effectively attend to the visual tokens and guide the PLM to classify sarcastic samples correctly. (4) Contrary to the general perception that multimodal methods should outperform unimodal ones, we find that this does not always hold true for few-shot scenarios. We hypothesize that in a multimodal scenario, the baseline models necessitate a larger parameter count for training, with only a limited amount of supervised data, which directly results in subpar performance. Our model CAMP outperforms the best multimodal baseline by 1.7% in MMSD and 0.7% in MMSD2.0 dataset. This is because CAMP only learns instance-specific continuous prompts while keeping the PLM frozen. Thus, CAMP can effectively utilize the knowledge base of the PLM while generating dynamic prompts that guide the PLM for better classification.

4.5 Out-of-Domain Evaluation

To assess the generalization ability of CAMP, we evaluate it on a new dataset, which we call **MCMD** (Multi-modal Code-Mixed Memes Dataset), introduced by (Maity et al., 2022). As there are only two publicly available multimodal sarcasm datasets, we opt for this dataset due to its similarity in nature and the presence of labeled sarcasm. To construct MCMD, we filter out memes without sarcasm labels or those that are code-mixed, resulting in 306 samples (183 sarcastic and 123 non-sarcastic). Since MMSD2.0 is a more balanced dataset, we train all models with it and test on MCMD.

It can be observed from Table 3 that our model shows a stronger generalization ability than other multimodal baselines[2] and methods in prompt-based finetuning strategy. We reason that since we don't change the PLM weights for CAMP during training, the PLM can retain its inherent knowledge of language understanding, which results in better performance for cross-dataset setup. We also observe that within the prompt tuning strategy, CAMP outperforms CAMP(w/o attn) because the continuous prompt vectors in CAMP can attend to the input modality tokens and thus can adapt to generate different continuous prompts based on the input instance.

5 Ablation Experiments

With our ablation experiments, we try to answer the following research questions. (1) *Is continuous attentive multimodal prompt better than its non-attentive counterpart?* (2) *How effective are continuous prompt tokens over their discrete counterparts for multimodal sarcasm detection?* (3) *Do captions reduce the semantic gap between image and text modalities?*

5.1 Attentive vs Non-Attentive

We evaluate CAMP and CAMP(w/o attn) on various continuous token lengths namely {10, 20, 30, 40, 50, 80, 100}. Figure 3 shows that accuracy increases for both models across both datasets as the

[2] HFM and HKE cannot be compared as they required external attributes which is not present for MCMD dataset.

		MMSD		MMSD2.0	
Strategy	**Method**	**Acc**	**F1**	**Acc**	**F1**
Prompting	PT_1^d	0.601	0.375	0.569	0.362
	PT_2^d	0.574	0.504	0.551	0.474
Prompt-Based	PT_1^d	0.735 (0.4)	0.722 (0.3)	**0.692** (1.6)	0.680 (1.6)
Finetuning	PT_2^d	**0.746** (0.9)	**0.731** (0.8)	0.688 (1.9)	0.687 (1.9)
Prompt	CAMP(w/o attn)	0.716 (0.5)	0.697 (0.7)	0.662 (0.2)	0.652 (0.4)
Tuning	CAMP	0.729 (0.9)	0.717 (1.0)	**0.692** (2.8)	**0.681** (2.3)

Table 5: Performance comparison of discrete vs continuous prompt-based methods. For Prompting approach, we only prompt the model on test set using the templates in Table 6. Hence, we do not report any standard deviation.

Discrete Prompt Templates	Label Words
$PT_1^d = [V_j]$ Tweet Text: $[T_j]$ Caption: $[C_j]$ Is the sentence sarcastic? $[MASK]$	Yes/No
$PT_2^d = [V_j]$ Tweet Text: $[T_j]$ Caption: $[C_j]$ The sentence is $[MASK]$	Sarcastic/Neutral

Table 6: Description of various discrete prompt templates that we design for ablation experiments. Here PT_i^d is the discrete prompt template i. Here $[V_j]$ stands for visual token slots, $[T_j]$ stands for textual token slots while C_j represents caption token slots.

number of tokens increases up to a certain point, after which the performance degrades. We reason that as prompt length increases, the PLM's ability to effectively capture the contextual nuances of the task at hand increases. However, after a certain point, the information learned by these tokens becomes redundant, which leads to overfitting. We find that CAMP performs superiorly over almost all continuous prompt token lengths than CAMP(w/o attn), with an average accuracy gain of +2.2% for MMSD and +1.33% for MMSD2.0 datasets. This shows the effectiveness of our attention module, which potently captures the dependencies between continuous tokens and the input tokens of text and image modalities.

5.2 Discrete vs Continuous

To demonstrate the effectiveness of continuous attentive multimodal prompt over its discrete counterparts, we formulate two discrete prompt templates, one in the declarative form and the other as an interrogative sentence, presented in Table 6. We also perform experiments with other templates and label words which are presented in Appendix section A.3. It can be seen from Table 5 that our proposed model CAMP which is based on prompt tuning strategy, outperforms prompting-based approaches by a very significant margin. This is because sarcastic utterances are less common in the general corpora on which these PLMs have been trained. We can also observe that a slight change in discrete prompts induces a significant difference in accuracy ($\triangle$2.7% for MMSD and $\triangle$1.8% for MMSD2.0) for prompting strategy. While prompt-based finetuning methods demonstrate a moderate performance advantage (+1.7% Acc in MMSD while no improvement in MMSD2.0) over our model, this outcome aligns with expectations, given that we do not finetune the entire PLM. Our model's strength lies in its parameter efficiency and consequently reduced training time, as we update only 30% of the entire model weights, compared to fine-tuning the entire model weights of the PLM.

5.3 Importance of Caption Tokens

The importance of caption tokens to bridge the semantic gap between image and text modalities can be seen from the reduced performance of the CAMP(w/o cap) variant in Table 4. This suggests that captions provide additional semantic information that enriches the context of an image. This additional layer of information helps the model better understand and interpret the image, leading to improved performance.

6 Conclusion

In this paper, we tackled the problem of few-shot multimodal sarcasm detection. Unlike traditional approaches that rely on early or late image-text fusion to learn the subtle interaction between the image and text modalities, we reformulate the problem as a cloze-filling task. To this end, we propose a novel approach of using continuous attentive multimodal prompt for this task. These attentive, continuous prompt tokens can effectively attend to the image and text modalities tokens and can dynamically adapt according to the input instance. Our extensive experiments over two datasets demonstrate the effectiveness of our model, which outperforms strong baselines in few-shot and Out-of-Distribution (OOD) settings. We also demonstrate the efficacy of our model CAMP over other discrete token-based techniques, including prompting and prompt-based finetuning, through several ablation experiments.

Limitations

Firstly, for our few-shot setting, we randomly sample 1% of the entire training dataset, which is an experimental choice. To account for the variability in sample diversity, we randomly sample two 1% splits of the training data and report the average performance. However, we believe that an alternate sampling strategy, in which more diverse samples can be collected, needs exploration. Secondly, some of the images have embedded text which we did not consider. Incorporating the text information present in the images could provide additional contextual cues and improve the overall understanding and analysis of the image content. For this study, we experimented with a BERT-base model. It will be interesting to see how other encoder or encoder-decoder architectures perform for the multimodal sarcasm detection task in the prompt-tuning paradigm.

References

Ameeta Agrawal and Aijun An. 2018. Affective representations for sarcasm detection. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2020. Leveraging transitions of emotions for sarcasm detection. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, Berlin, Germany.

Joshua M. Averbeck. 2013. Comparisons of ironic and sarcastic arguments in terms of appropriateness and effectiveness in personal relationships. *Argumentation and Advocacy*, 50:47 – 57.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online).

Rohan Badlani, Nishit Asnani, and Manan Rai. 2019. An ensemble of humour, sarcasm, and hate speechfor sentiment classification in online reviews. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, Hong Kong, China. Association for Computational Linguistics.

Andrew Brock, Soham De, and Samuel L. Smith. 2021. Characterizing signal propagation to close the performance gap in unnormalized resnets. *ArXiv*, abs/2101.08692.

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*, Florence, Italy. Association for Computational Linguistics.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. *Proceedings of the 31st ACM International Conference on Multimedia*.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christine Chappuis, Valérie Zermatten, Sylvain Lobry, B. L. Saux, and Devis Tuia. 2022. Prompt–rsvqa: Prompting visual context to a language model for remote sensing visual question answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, Uppsala, Sweden. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the 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. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Megan L. Dress, Roger J. Kreuz, Kristen E. Link, and Gina M. Caucci. 2008. Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27:71 – 85.

Simona Frenda. 2018. The role of sarcasm in hate speech.a multilingual perspective.

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)*, Online. Association for Computational Linguistics.

José Antonio García-Díaz, Ronghao Pan, and Rafael Valencia-García. 2023. Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english. *Mathematics*.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Saarbrücken, Germany.

Debanjan Ghosh, Ritvik Shrivastava, and Smaranda Muresan. 2021. "laughing at you or with you": The role of sarcasm in shaping the disagreement space. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online. Association for Computational Linguistics.

Roberto I. González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Annual Meeting of the Association for Computational Linguistics*.

Alex Graves and Jürgen Schmidhuber. 2005. 2005 special issue: Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven C. H. Hoi. 2022. From images to textual prompts: Zero-shot visual question answering with frozen large language models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States. Association for Computational Linguistics.

Stacey L. Ivanko and Penny M. Pexman. 2003. Context incongruity and irony processing. *Discourse Processes*, 35:241 – 279.

Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying creative harmful memes via prompt based approach. *Proceedings of the ACM Web Conference 2023*.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Contrastive learning for prompt-based few-shot language learners. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China. Association for Computational Linguistics.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas.

Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. 2015. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Lisboa, Portugal.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *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)*, abs/2101.00190.

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. *Proceedings of the 29th ACM International Conference on Multimedia*.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multimodal sarcasm detection via cross-modal graph convolutional network. In *Annual Meeting of the 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*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *ArXiv*, abs/2103.10385.

Yiyi Liu, Yequan Wang, Aixin Sun, Xuying Meng, Jing Li, and Jiafeng Guo. 2022b. A dual-channel framework for sarcasm recognition by detecting sentiment conflict. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States.

Yuhang Liu, Wei Wei, Daowan Peng, and Feida Zhu. 2022c. Declaration-based prompt tuning for visual question answering. In *International Joint Conference on Artificial Intelligence*.

Chenwei Lou, Bin Liang, Lin Gui, Yulan He, Yixue Dang, and Ruifeng Xu. 2021. Affective dependency graph for sarcasm detection. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Timothy Ossowski and Junjie Hu. 2023. Multimodal prompt retrieval for generative visual question answering. *ArXiv*, abs/2306.17675.

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

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, Hong Kong, China. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. MMSD2.0: Towards a reliable multimodal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Patricia Rockwell and Evelyn M. Theriot. 2001. Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, 18:44 – 52.

Rossano Schifanella, Paloma de Juan, Joel R. Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. *Proceedings of the 24th ACM international conference on Multimedia*.

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)*, Toronto, Canada. Association for Computational Linguistics.

Christopher W. Tindale and James Gough. 1987. The use of irony in argumentation. *Philosophy and Rhetoric*, 20:1–17.

Frans H. van Eemeren and Rob Grootendorst. 1992. Argumentation, communication, and fallacies: A pragma-dialectical perspective.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix X. Yu, Cho-Jui Hsieh, Inderjit S. Dhillon, and Sanjiv Kumar. 2022. Two-stage llm fine-tuning with less specialization and more generalization.

Chan Shao Wen, Guoli Jia, and Jufeng Yang. 2023. Dip: Dual incongruity perceiving network for sarcasm detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*, Online. Association for Computational Linguistics.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songiln Hu. 2022a. Multimodal hate speech detection via cross-domain knowledge transfer. *Proceedings of the 30th ACM International Conference on Multimedia.*

Xiaocui Yang, Shi Feng, Daling Wang, Pengfei Hong, and Soujanya Poria. 2022b. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. *Proceedings of the 31st ACM International Conference on Multimedia.*

Yang Yu and Dong Zhang. 2022. Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. *2022 IEEE International Conference on Multimedia and Expo (ICME).*

Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. *Proceedings of the 30th ACM International Conference on Multimedia.*

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Y. Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *ArXiv*, abs/2309.10313.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan.

A Appendix

A.1 Hyperparameter Details

We run all our experiments on a Nvidia RTX A5000 GPU with 24GB of memory. We use the pre-trained blip-opt-2.7b [3] model for generating captions. We employ the OpenPrompt[4] library to build our prompt learning model. All our experiments use AdamW optimizer with a weight decay of 0.01. We run our model for 20-100 epochs and pick the model that performs best for the validation set for testing. In all experiments, we use a learning rate of 0.0001 and a batch size of 16. The value of h_{dim} is 768, which is the default embedding dimension for BERT. The number of continuous prompt tokens is set to 50 for MMSD and 80 for MMSD2.0, while image token slots are fixed at 3 for both datasets. The maximum token length for the PLM is 128.

A.2 Performance on Different Discrete Tokens

In this section, we experiment with different discrete tokens shown in Table 7 and present their comparative analysis in Table 10 in both prompting and prompt-based finetuning techniques. Sarcasm detection, being a difficult task, simple prompting with discrete tokens yields sub-optimal performance while showing a lot of variation in performance. However, finetuning the entire parameter set of BERT demonstrates a significant jump in performance, which is expected.

Discrete Prompt Templates	Label Words
$PT_1^d = [V_j]$ Tweet Text: $[T_j]$ Caption: $[C_j]$ Is the sentence positive? $[MASK]$	Yes/No
$PT_2^d = [V_j]$ Tweet Text: $[T_j]$ Caption: $[C_j]$ So the meme is: $[MASK]$	Sarcastic/Neutral
$PT_3^d = [V_j]$ Tweet Text: $[T_j]$ Caption: $[C_j]$ This post can be termed as: $[MASK]$	Funny/Serious

Table 7: Description of Various Discrete Prompt Templates. Here PT_i^d is the discrete prompt template i. Here $[V_j]$ stands for visual token slots, $[T_j]$ stands for textual token slots while C_j represents caption token slots.

	MMSD	
Method	**Acc**	**F1**
ResNet	0.715	0.696
ViT	0.663	0.659
NF-ResNet-50	**0.729**	**0.717**

Table 8: Performance comparison of different visual encoders for our CAMP model.

A.3 Effect of Different Visual Encoders

We experimented with different visual encoders, including ResNet and ViT, for our CAMP model. The experimental results on MMSD dataset are presented in Table 8. However, we found NF-ResNet-50 performs the best among them and hence we use this for all our experiments.

	MMSD	
Image Token Length	**Acc**	**F1**
1	0.710	0.671
3	**0.729**	**0.717**
5	0.724	0.702
7	0.716	0.699

Table 9: Ablation experiment on different image tokens for CAMP.

A.4 Impact of Different Image Token Lengths

To find out how much image information is required for CAMP to achieve best performance, we conduct experiments with varied image token lengths on MMSD dataset. The length of continuous prompt token is kept at 50 since we achieve best performance for MMSD dataset. It can be observed from Table 9 that the when image token

[3] https://huggingface.co/Salesforce/blip2-opt-2.7b

[4] https://github.com/thunlp/OpenPrompt

		MMSD		**MMSD2.0**	
Strategy	**Method**	**Acc**	**F1**	**Acc**	**F1**
Prompting	PT_1^d	0.601	0.377	0.567	0.364
	PT_2^d	0.603	0.501	0.554	0.497
	PT_3^d	0.436	0.435	0.491	0.491
Prompt-Based Finetuning	PT_1^d	0.732 (0.1)	0.727 (0.4)	0.686 (0.1)	0.664 (0.1)
	PT_2^d	0.721 (0.5)	0.718 (0.6)	0.702 (0.8)	0.691 (0.1)
	PT_3^d	0.738 (0.8)	0.718 (0.3)	0.694 (2.4)	0.691 (2.4)

Table 10: Performance comparison of discrete prompts under prompting and prompt-based finetuning strategy. Numbers in bracket indicate standard deviation. For Prompting approach, we only prompt the model on test set using the templates in Table 7. Hence, we do not report any standard deviation.

length is 1, the utilization of image information becomes incomplete, whereas increasing it beyond 3 introduces redundancy to the model.