Meta Review of Submission202 by Area Chair Hef8

Meta Review by Area Chair Hef8 06 Dec 2024, 18:55 (modified: 13 Dec 2024, 17:24) Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs Revisions

Metareview:

The paper presents a visually-grounded prompt tuning approach to improve the task of categorical emotion detection (CED), titled Visual Prefix-guided Emotion Detector (VISPOR). This involves aligning textual descriptions with related images in a shared semantic space (based on the pre-trained CLIP model), which is used to generate visual-enriched prefixes for transformer models. The authors evaluate their method on three CED benchmarks (ISEAR, TEC and CARER), stating achieved state-of-the-art results over existing methods across all three benchmarks.

Although the weaknesses identified by the reviewers do not seem to be major, the overall assessment scores are rather low (2.5, 2.5 and 3.5).

Summary Of Reasons To Publish:

- The authors propose an interesting and reasonable approach to CED.
- Evaluation on multiple datasets and comparison to various baselines showcases the robustness of the method.
- Justification for the method is provided through ablation studies.

Summary Of Suggested Revisions:

Based on the reviewers' comments, the study does not require additional experiments, but the paper should be revised to make some things clearer:

- Add details about the baseline models.
- Add details about some parts of the methodology (e.g. processing of the datasets, explanation of general prefix; see comments by individual reviewers).
- Situate the work in the broader context of multimodal emotion detection (e.g. work on IEMOCAP dataset).
- Extend limitations section with discussion on cultural bias and adaptability to other models (cf. fact that the method depends heavily on the pre-trained CLIP model).

Overall Assessment: 4 = There are minor points that may be revised
Best Paper Ae: No
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Author Identity Guess: 1 = I do not have even an educated guess about author identity.
Reported Issues: No

Add: **Author-Editor Confidential Comment**

The more, the merrier: Detecting Categorical Emotions from Texts with Cross-modal Insights

Anonymous ACL submission

Abstract

Categorical Emotion Detection (CED) task aims to identify the emotion expressed in a given text. While Prompt Tuning (PT) has been applied to the CED, existing detectors struggle to design task-efficient prompts. Given that human perception is cross-modal, we propose a **Vis**ual Emotion **P**refix-guided Emotion Detect**or** (VISPOR) that exploits visual information from emotion resources to heuristically construct prompts that better adapt to the CED. Notably, to mitigate the severe modality gap, the VISPOR aligns abstract textual descriptions of emotions with concrete emotional images within a shared space. Then, by taking such visual-enriched text embeddings as prefix, the visual emotion information can smoothly and effectively serve the VISPOR to capture nuanced emotion features from texts. We conduct extensive experiments on three CED benchmarks, and the results show that VISPOR significantly outperforms existing methods.

1 Introduction

Human emotions significantly shape thoughts, behaviors, and decisions, exerting a profound influence on social interactions and relationships. Textual data serves as a critical medium through which individuals express their emotional states (Deng and Ren, 2023). Additionally, with the rapid development of social media platforms such as Twitter and Reddit, people are able to communicate and express their opinions more freely than ever before, providing a rich foundation for data-driven, emotion-centric research. However, some of these opinions may be illegal or extreme, often carrying negative emotions. Thus, understanding and detecting emotions from large-scale textual data is of paramount importance.

During the past decades, the Categorical Emotion Detection (CED) task has been widely investigated by the community, suggesting a number of traditional emotion detectors and the emerging deep neural network (DNN)-based ones. Earlier emotion detectors mainly capture word-level emotion clues through affective resources, such as emotion lexicons (Mohammad and Turney, 2013), to detect emotion categories (Bao et al., 2011; Katz et al., 2007). Subsequent topic-driven approaches (Wang et al., 2019) assume that samples within the same emotion category may share similar topic distributions, thus learning complementary emotion semantics by incorporating either statistical or neural topic models (NTM). However, the limited scales of training samples may hinder the accuracy of topic distribution estimation.

More recently, Prompt Tuning (PT) has emerged as a novel fine-tuning paradigm, achieving promising results across various NLP tasks (Hu et al., 2022; Han et al., 2022; Liu et al., 2023). It constructs predefined templates or virtual vectors as prompts, updating only these parameters to improve the task performance. Though this paradigm has also been explored in the context of the CED task (Plaza-del Arco et al., 2022; Bareiß et al., 2024), their ways of designing prompts are still inefficient.

We notice that human perception is internally cross-modal. *When perceiving emotions through text, the same neurons are activated in the cerebral cortex as when emotions are perceived through other modalities* (Gibson, 1969; Meltzoff and Borton, 1979; Lin et al., 2023). For example, anger texts may evoke the same neuron response as visual images depicting anger scenarios. This finding motivates us to acquire sufficient information from emotion images, enriching the emotion-specific prompts with the cross-modal augmentation.

However, directly encoding emotion images into the prompts and applying to emotion texts may bring the modality gap (Liang et al., 2022), compromising the task performance. To address this issue, we jointly exploit emotion images so-called

concrete concepts and class-wise text descriptions of emotions to construct a shared semantic space, due to humans' emotion understanding results from the interaction of both the abstract and concrete concepts (Lakoff and Johnson, 2008). With sufficient interactions between the two modalities, the visual information can be well injected into these text descriptions of emotions. We then formulate their embeddings as complementary visual-enriched prompts, which better adapt to the CED task.

Upon these ideas, we propose a novel CED method entitled **Vis**ual **P**refix-guided Emotion Detec**tor** (VISPOR), which operates in two stages: cross-modal prefix generation and visual-guided prefix tuning. The first stage generates a cross-modal prefix with auxiliary visual information. Specifically, we set a text description for each emotion category and refine quality emotion images from datasets by calculating class-wise text-image similarity with a pre-trained CLIP (Radford et al., 2021). Then, these selected images are paired with the text descriptions. We exploit such manually-aligned pairs to derive a shared semantic space via fine-tuning the CLIP. The encoded text descriptions carried rich visual information are taken as the cross-modal emotion prefix, which can naturally adapt to the CED task. In the second stage, the obtained prefix is then injected into each self-attention layer of a Transformer-based emotion detector. Eventually, we train the VISPOR by only optimizing parameters of the prefix and an additional classifier with CED samples, resulting in better emotion predictions. In a nut shell, the contributions of our work are as follows:

- We present the VISPOR that resolves the CED task by exploiting rich visual information in the PT paradigm, as human perception is internally cross-modal.
- We align text descriptions of emotion categories with the corresponding images in a shared semantic space, alleviating the modality gap.
- We evaluate the proposed VISPOR over three CED benchmarks. The experimental results show that our method consistently performs better than the existing state-of-the-art (SOTA) baselines.

2 Related Work

In this section, we mainly revise current researches on PT and CED.

Prompt Tuning. We first deliver an overview of current prompt tuning (PT) methods in the NLP community. The basic idea of PT is to exploit natural-language prompts and task demonstrations as context to make downstream tasks close to language modeling (Brown, 2020). Early studies manually construct templates so-called hard templates, to improve text classification and natural language inference tasks (Schick and Schütze, 2021; Hu et al., 2022). However, designing these task-oriented templates requires strong domain knowledge. As a result, some automatic generation methods for hard templates are explored (Wu et al., 2022). Hard prompts have good interpretability but lack generalizability, making them difficult to directly transfer to other complex tasks.

More recently, Li and Liang (2021) propose the prefix-tuning paradigm that only optimizes a small, task-specific sequence of continuous vectors while keeping the core language model frozen. This paradigm has been widely adopted to a series of NLP tasks, such as event extraction (Liu et al., 2022), fine-grained sentiment analysis (Wu and Shi, 2022), and text classification (Wen and Fang, 2024). Here, we exploit rich information from emotion images to construct visual-enriched emotion prefix, improving the CED task in the prefix-tuning paradigm.

Categorical Emotion Detection. Earlier CED approaches (Bao et al., 2011; Katz et al., 2007; Lei et al., 2014) mainly capture latent features from individual words, further applying machine learning algorithms (Cambria et al., 2015) to achieve classification. For instance, Katz et al. (2007) construct a word-emotion mapping dictionary to score the emotion intensity inherent in headlines. However, such word-based methods fail to accurately perceive emotions carried by a same word with different contexts.

With a breakthrough of DNN, the neural-induced emotion detectors mainly capture deep emotion semantics with auxiliary topic distributions, which estimated from either statistical or neural topic models (Cao et al., 2015; Gui et al., 2019). For instance, Wang et al. (2019) simultaneously exploit external topic distributions and syntactic structures, learning quality emotion features guided by topic clusters. Later, TESAN (Wang and Wang, 2020)

Table 1: Summary of Notations.

Notation	Description
N	number of CED samples
C	number of emotion categories
$\mathcal{D}_p$	the paired text-image dataset
$\mathbf{y} \in \{0,1\}^C$	emotion label space
$\mathbf{E}$	the set of emotion descriptions
Θ	parameters of CLIP text encoder
Φ	parameters of CLIP visual encoder
$\mathcal{O}_{\text{Prx}}$	the CED prefix

learns topic-enriched emotion semantics by jointly training the emotion detector with an auxiliary neural topic model. Although achieving remarkable performances, the quality of the derived topic distributions heavily depend on the size of training samples. More recently, a number of studies (Plaza-del Arco et al., 2022; Bareiß et al., 2024) investigate the effectiveness of the PT paradigm to train the emotion detectors. However, these attempts directly transfer general PT frameworks to the CED task, ignoring the inherent nuances across various emotions.

Orthogonal to the aforementioned emotion detectors, we utilize emotion text descriptions to guide the VISPOR in better absorbing rich information from emotion images, thereby constructing a visual-enriched prefix from the cross-modal learning perspective.

3 Methodology

In this section, we briefly describe the task definition of CED. Then, we introduce the proposed VISPOR as well as its components in more detail. For clarity, some important notations are summarized in Table 1.

Task definition. Given N labeled training samples, the CED task aims to induce an emotion detector enabling to identity which emotions each text sample contains. Formally, each training sample is represented by (s_i, y_i), where $s_i = \{w_{i1}, \cdots, w_{iM}\}$ is the raw text and $y_i \in \mathcal{Y}$ is the category label. In our work, the label space is commonly defined by following the Ekman emotion theory [1].

3.1 Overview of VISPOR.

As depicted in Fig.1, our VISPOR mainly operates in two stages: (1) **Cross-modal Prefix generation**: We design a set of text descriptions $\mathbf{E} = \{\mathbf{e}_0, \mathbf{e}_1, \cdots, \mathbf{e}_C\}$ for each emotion. Then, we calculate class-wise similarity between $\mathbf{e}_i \in \mathbf{E}$ and emotion images $\mathcal{D} = \{\mathbf{d}_0, \mathbf{d}_1, \cdots, \mathbf{d}_{N_p}\}$, keeping the top M_p most-related ones. We pair these selected images with the corresponding emotion descriptions, constructing an aligned text-image dataset $\mathcal{D}_p = \{\mathbf{e}_i, \mathbf{d}_i\}_{i=1}^{M_p \times C}$. We fine-tune the CLIP with the $\mathcal{D}_p$ to construct a shared semantic space where visual information of emotions can be fully injected into the text descriptions $\mathbf{E}$. Their embeddings $\mathbf{H} = [\mathbf{h}_0, \ldots, \mathbf{h}_{C-1}]^\top$ are considered as the visual-enriched emotion prefix.

(2) **Visual-guided Prefix Tuning**: We concatenate the obtained emotion prefix $\mathbf{H}$ with a general prefix $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_{k-1}]^\top$ to formulate the final CED prefix $\mathcal{O}_{\text{Prx}}$. We then suggest a PLM and incorporate the $\mathcal{O}_{\text{Prx}}$ in each layer. By training with prefix-tuning, the PLM derives a visual-enriched emotion representation $\mathbf{z}_{i\,[\text{CLS}]}^{(L)}$ for each CED text. We predict its emotion label $\hat{y}_i$ with a single-layer MLP. In the following, we introduce each component of VISPOR in more detail.

3.2 Cross-modal Prefix Generation.

Paired Text-image Construction. We first construct a set of text-image pairs for the CED, aiming to establish a shared semantic space where visual and textual modalities can be fully interacted. Specifically, we set a list of text descriptions for each emotion state $\mathbf{E} = \{\mathbf{e}_0, \mathbf{e}_1, \cdots, \mathbf{e}_C\}$, each of which is constructed with a template of "*This image expresses [EMOTION STATE]*".

Meanwhile, we collect a set of emotion images[2] $\mathcal{D} = \{\mathbf{d}_0, \mathbf{d}_1, \cdots, \mathbf{d}_{N_p}\}$, where N_p is the total number of images. For each emotion category, we utilize a pre-trained CLIP to compute the similarity between the images and the corresponding emotion description. Given an emotion image $\mathbf{d}_i \in \mathcal{D}$ and its text description $\mathbf{e}_j$, its similarity score t_i is calculated as follows:

$$t_i = \text{CLIP}(x_i, e_j) \tag{1}$$

For the j-th emotion category, we rank the images based on the similarity scores $T^{(j)} = \{t_0, t_1, \cdots, t_{N_p^{(j)}}\}$, where $N_p^{(j)}$ denotes the number of images belong to the j-th emotion category, in descending order and select the top-M_p images to form a subset. This subset, along with the $\mathbf{e}_j$,

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

[2] These images are collected from several emotion image datasets, including EmoSet (Yang et al., 2023), FlickrLDL (Yang et al., 2017), and EMOTIC (Kosti et al., 2017).

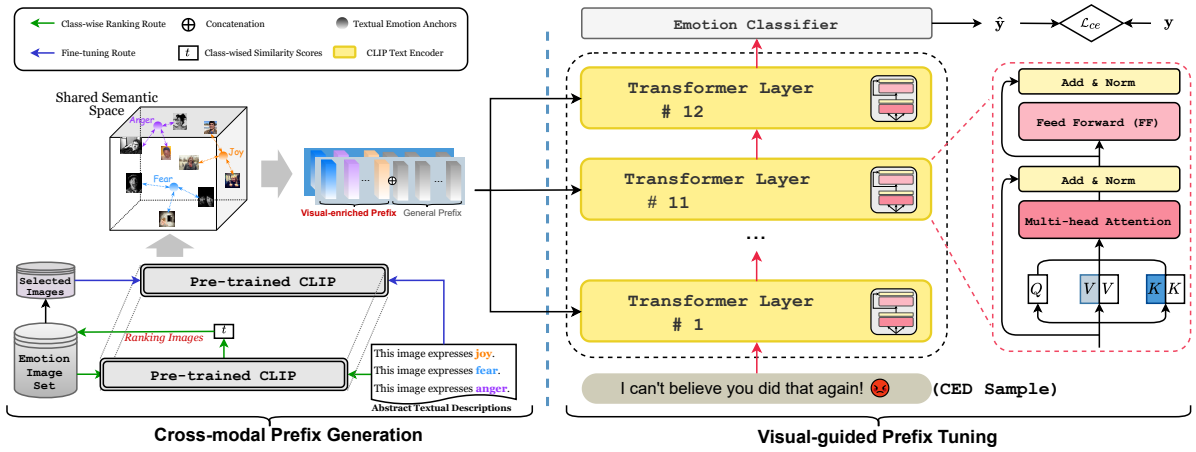

Figure 1: The overall framework of the VISPOR.

constitutes a set of text-image pairs denoted as $\mathcal{D}^{(j)} = \{\mathbf{e}_j, \mathbf{d}_i^{(j)}\}_{i=1}^{M_p}$. The comprehensive set of text-image pairs across all emotions is represented as $\mathcal{D}_p = \bigcup_{j=1}^{C} \mathcal{D}^{(j)}$.

CLIP Fine-tuning. Based on the aligned multimodal dataset $\mathcal{D}_p$, we then fine-tune the CLIP, which consists of an image encoder CLIP_{vis} and a transformer text encoder CLIP_{txt}. Given a text-image pair $(\mathbf{e}_i, \mathbf{d}_i) \in \mathcal{D}_p$, the two encoders respectively learn the embeddings of the two modalities as follows:

$$u_i = \text{CLIP}_{\text{txt}}(e_i; \Theta) \tag{2}$$

$$v_i = \text{CLIP}_{\text{vis}}(x_i; \Phi) \tag{3}$$

where Θ and Φ denote the trainable parameters of the two encoders. Then, the CLIP is fine-tuned with contrastive losses based on the obtained embeddings u and v:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}(\mathcal{L}_{vis \to txt} + \mathcal{L}_{txt \to vis}) \tag{4}$$

and the $\mathcal{L}_{vis \to txt}$ (similar to $\mathcal{L}_{txt \to vis}$) is formulated as:

$$\mathcal{L}_{vis \to txt} = -log \frac{\exp(\text{sim}(v_i, u_i)/\tau)}{\exp(\Sigma_{j=0}^{N_s} \text{sim}(v_i, u_j)/\tau)} \tag{5}$$

where N_s denotes the number of samples in the mini-batch, sim is the cosine similarity, and τ is the temperature.

By optimizing the $\mathcal{L}_{\text{CLIP}}$, we obtain a shared semantic space in which each text description are closely aligned with the emotion images from the same emotion category. In this space, the embeddings $\mathbf{H} = [\mathbf{h}_0, \ldots, \mathbf{h}_{C-1}]^\top$ of the text descriptions have carried sufficient visual information of emotions and are taken as an emotion prefix in the subsequent tuning stage. This way enables the auxiliary visual information to more effectively support the CED task, while reducing the detrimental effects caused by the modality gap.

3.3 Visual-guided Prefix Tuning.

After obtaining the emotion prefix $\mathbf{H}$, we set a PLM and generates visual-enriched emotion predictions for each CED text with prefix-tuning.This PLM share parameters with the well-trained CLIP_{txt}. Considering that the text contains both emotion-relevant and emotion-irrelevant parts, we concatenate the $\mathbf{H}$ with a set of general prefix vectors $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_{k-1}]^\top$, formulating the final CED prefix $\mathcal{O}_{\text{Prx}} = [\mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_{C-1}, \mathbf{b}_0, \mathbf{b}_1, \cdots, \mathbf{b}_{k-1}]^\top$.

We then prepend this synthetic prefix $\mathcal{O}_{\text{Prx}}$ to the text sequence at each self-attention layer of the Transformer-based encoder of the PLM. Formally, given a CED text $s_i = \{w_{i1}, \cdots, w_{iM}\}$, its contextual embeddings $\mathbf{Z}^{(l-1)} = \{\mathbf{z}_0^{(l-1)}, \mathbf{z}_1^{(l-1)}, \cdots, \mathbf{z}_M^{(l-1)}\}$ are firstly respectively mapped into *Query*, *Key*, and *Value* vectors at the l-th layer as follows:

$$\mathbf{Q}^l = \mathbf{W}_Q^l \mathbf{Z}^{(l-1)} \tag{6}$$

$$\mathbf{K}^l = \mathbf{W}_K^l \mathbf{Z}^{(l-1)} \tag{7}$$

$$\mathbf{V}^l = \mathbf{W}_V^l \mathbf{Z}^{(l-1)} \tag{8}$$

where $\mathbf{W}_Q^l$, $\mathbf{W}_K^l$, and $\mathbf{W}_V^l$ denote trainable parameters of the l-th layer.

Formally, the CED prefix-guided attention of the l-th layer are calculated as follows:

$$\text{Attn}^l = \sigma\left(\frac{\boldsymbol{Q}^l \left[\mathcal{O}_{\text{Prx}}^l; \boldsymbol{K}^l\right]^\top}{\sqrt{d}}\right)\left[\mathcal{O}_{\text{Prx}}^l; \boldsymbol{V}^l\right] \tag{9}$$

Table 2: Statistics and data splitting of the CED datasets. # denotes the number of samples.

Datasets	#Train	#Dev	#Test	Emotion Labels
ISEAR	5,366	767	1,533	*anger, disgust, fear, joy, sadness, shame, guilt*
CARER	14,000	2,000	4,000	*anger, love, fear, joy, sadness, surprise*
TEC	14,736	2,105	4,210	*anger, disgust, fear, happy, sadness, surprise*

where d denotes the dimension of the textual embeddings. By sequentially conducting prefix-guided attention to update all textual states, the CED texts can be fully interacted with the rich visual information of emotions. Then, the final textual states encode both the context and the cross-modal emotion semantics simultaneously.

3.4 Classifier and Training Objective.

We leverage embeddings of the **[CLS]** token to represent the ingested texts and obtain the final sample embeddings $\{\mathbf{z}_{i\,[\mathrm{CLS}]}^{(L)}\}_{i=1}^{N}$ from the PLM. Then, we employ a single-layer MLP as the emotion classifier. For each $\mathbf{z}_{i\,[\mathrm{CLS}]}^{(L)}$, we predict its category label $\hat{y}_i$ by the following equation:

$$\hat{y}_i = \mathrm{Softmax}\left(\mathbf{W}_c \mathbf{z}_{i\,[\mathrm{CLS}]}^{(L)}\right), \qquad (10)$$

where $\mathbf{W}_c$ is the trainable parameter of the emotion classifier.

Regard N training pairs $\{(\mathbf{z}_{i\,[\mathrm{CLS}]}^{(L)}, y_i)\}_{i=1}^{N}$, we then formulate the final objective of VISPOR with respect to all trainable parameters $\mathbf{W} = \{\mathcal{O}_{\mathrm{Prx}}, \mathbf{W}_c\}$:

$$\mathcal{L}(\mathbf{W}) = \sum_{i=1}^{N} \mathcal{L}_{\mathrm{CE}}(y_i, \hat{y}_i) + \lambda \|\mathbf{W}\|^2, \qquad (11)$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss (Gneiting and Raftery, 2007); $\|\cdot\|$ denotes the ℓ_2-norm; and $\lambda \in [0, 1]$ is the regularization coefficient.

4 Experiments

4.1 Experimental Settings.

Benchmarks. We evaluate the VISPOR and the compared CED methods on three public available datasets in English. The details of the datasets are shown in the Table 2.

- ISEAR dataset contains 7,666 samples that manually annotated by about 1,000 participants into seven emotions.
- TEC dataset (Mohammad, 2012) contains 21,051 tweets from about 19,000 different users. These tweets are collected from Twitter via searching with hashtags corresponding to the six emotions.
- CARER dataset (Saravia et al., 2018) contains 20,000 English tweets filtered by 339 hashtags from Twitter API, and the hashtag appearing in the last position of a tweet is treated as the ground truth. These samples are categorized into six emotions.

Baselines. We select the following CED methods that are either traditional or neural-based for comparison.

ETM (Bao et al., 2011): A topic-based method that introduces an extra emotion layer to the LDA.

DACNN (Yang and Chen, 2020): An emotion detector that applies a multi-channel convolutional network with attention to improve emotion categorization performance.

WED (Li et al., 2021c): A knowledge-driven emotion detector that exploits domain knowledge and lexicons to generate an affective representation of a word using fine-grained emotion concepts.

EmoChannel-SA (Li et al., 2021b): An emotion detector that exploits fine-grained emotion knowledge sourced from dimensional sentiment lexicons with self-attention mechanism.

AGN (Li et al., 2021a): An emotion detector that proposes an Adaptive Gate Network to consolidate semantic representation by selectively fusing the corpus-level and word-to-label features.

Gated DR-G-T (Wang et al., 2019): An emotion detector enriched by syntactic structures and statistical topic distributions. A 300-dimensional word2vec model [3] is employed to initialize texts, and the topic number of the Latent Dirichlet Allocation (LDA) is set to 10.

TESAN (Wang and Wang, 2020): An end-to-end emotion detector with a neural topic model (NTM) that jointly learning topic embeddings and

[3] https://code.google.com/archive/p/word2vec/

Table 3: The experimental results of all comparing methods in terms of Accuracy (Acc) and Macro-F1 (F1). The best results are represented in **bold**. The second-best results are *underlined*.

Datasets	ISEAR		TEC		CARER	
Metric	Acc	F1	Acc	F1	Acc	F1
ETM	50.49	48.66	43.91	34.77	56.47	49.21
Bi-LSTM	55.27	54.59	54.76	43.85	69.58	62.94
TextCNN	57.42	56.71	55.02	44.58	69.24	62.37
BERT	61.57	60.80	60.14	51.42	72.70	65.09
DACNN	58.28	56.94	57.57	48.10	70.85	63.31
WED	58.74	58.02	57.24	47.83	71.16	64.32
EmoChannel-SA	59.09	58.45	58.52	49.15	72.03	65.38
AGN	59.28	58.33	58.97	50.66	71.55	64.38
Gated DR-G-T	59.13	58.86	57.64	47.95	68.63	61.80
TESAN	58.62	57.93	57.94	48.05	70.58	63.39
STN	60.83	59.44	59.18	50.04	73.07	65.88
VISPOR	**63.72**	**62.18**	**62.69**	**53.30**	**75.65**	**68.24**

document embeddings via attention mechanism. The topic number of the NTM is set to 30.

STN (Dai et al., 2022): An fast training emotion detector that sufficiently exploits rich semantic inherent in each topic. The topic number of the LDA is set to 10.

Moreover, we also compare with several commonly-used text classification methods, *i.e.,* Bi-LSTM, TextCNN, and BERT [4].

Training Details. All the neural-based methods are implemented in PyTorch and run with a NVIDIA TITAN RTX GPU. Specifically, the model is trained by Adam optimizer with a learning rate of 1e-5. The length of the general prefix k is set to 6. The number of selected emotion images M_p for each category is set to 100. The number of epoch is set to 5. The batch size is set to 32. We take Accuracy (Acc) and Macro-F1 (F1) as the metrics. In terms of all datasets, the splitting of training and testing sets is shown in Table 2. We take the average of the results from five independent runs and consider it as the final results.

4.2 Results and Analysis.

Table 3 displays the comparative performance of the baseline emotion detectors and the proposed VISPOR. Across all evaluation metrics, our VISPOR shows consistent and substantial improvements over these baselines. To be specific, compared to the existing SOTA method BERT, VISPOR achieves 1.38 and 1.88 improvements in terms of the F1 scores across ISEAR and TEC, respectively. The results of the BERT are re-produced through fine-tuning on the training set, where all parameters are updated. Nevertheless, the VISPOR can still surpass it by updating less than 1% of the parameters. This indicates that exploiting rich information from emotion images in the joint text-image semantic space via prefix-tuning can effectively enhance the task performance of CED.

In addition, the CARER is an imbalanced dataset, yet the VISPOR can still achieve prominent performances under the imbalance condition. Turning to the comparison of neural-based detectors, we observe that detectors driven by external resources, such as STN, EmoChannel-SA, and BERT, generally outperform those without such resources, including Bi-LSTM, TextCNN, and DACNN. This suggests that, for both humans and machines, understanding emotions relies on the external knowledge. Moreover, across the majority of evaluation metrics, topic-based detectors *i.e.,* Gated DR-G-T, TESAN, and STN, are superior to ones based on word distributions. This may be because the derived topic clusters can group certain emotion-laden words.

[4] https://huggingface.co/bert-base-uncased

Table 4: The ablation results of VISPOR in terms of Accuracy (Acc) and Macro-F1 (F1). The best results are represented in **bold**. The sub-optimal results are *underlined*. **w/o** and **r/p** respectively denote "without" and "replace".

Datasets	ISEAR		TEC		CARER	
Metric	Acc	F1	Acc	F1	Acc	F1
w/o Images	61.09 ↓	60.21 ↓	61.48 ↓	51.72 ↓	74.20 ↓	66.82 ↓
w/o Texts	61.77 ↓	60.53 ↓	62.15 ↓	51.78 ↓	74.46 ↓	66.75 ↓
w/o EP	60.31 ↓	58.95 ↓	59.69 ↓	51.14 ↓	73.39 ↓	65.74 ↓
r/p BERT	62.15 ↓	61.59 ↓	61.32 ↓	52.50 ↓	73.84 ↓	66.26 ↓
r/p RoBerta	62.82 ↓	61.37 ↓	**62.75** ↑	53.26 ↓	74.98 ↓	67.73 ↓
VISPOR	**63.72**	**62.18**	62.69	**53.30**	**75.65**	**68.24**

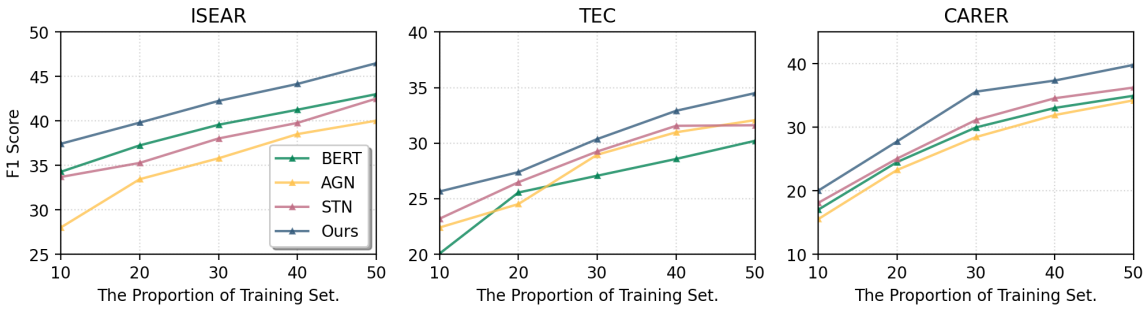

Figure 2: F1 scores of the VISPOR and compared emotion detectors in the low-source setting.

4.3 Ablation Study.

Moreover, we conduct ablative experiments to evaluate the effectiveness of the important components of the VISPOR below: *w/o Images:* removing the images and only using the text descriptions **E** encoded by the CLIP as the emotion prefix; *w/o Texts:* removing the text descriptions and averaging the image representations of each category as the emotion prefix; *w/o EP:* removing the emotion prefix **H** and only using the general prefix. The results are shown in Table 4.

From the results, we can observe that the performance of the VISPOR drops significantly across all three datasets when removing text descriptions. This may be due to the absence of guidance from abstract text descriptions, leading to a substantial modality gap when directly encoding emotion images as the prefix. Besides, we observe that removing emotion images has a greater impact on task performances than removing the text descriptions, suggesting that emotion images so-called concrete concepts contribute more to understanding emotions than those abstract emotion concepts. Moreover, when the EP is removed, the performance of the VISPOR deteriorates the most. This validates the importance of emotion information from the visual modality to the CED task, aligning with the statement that human perception is internally cross-modal.

In addition, we replace the text encoder with two commonly-used PLMs[5]: *r/p BERT* and *r/p RoBerta*. We observe that after the replacements, the performances of the VISPOR generally decline in most cases. This indicates that leveraging the CLIP text encoder that simultaneously absorbs both textual and visual semantics can more effectively exploit the visual-enriched emotion prefix. However, we find that on the TEC dataset, replacing with RoBerta leads to a slight improvement in terms of Acc. This may be because, when fine-tuning the CLIP, the images may not be fully aligned with the text descriptions.

4.4 Low-source Scenario.

We further conduct experiments in low-resource scenario by randomly sampling 10% to 50% from the original training set to form a low-resource training set. Fig.2 shows the performance of the VISPOR in a low-resource scenario compared with several powerful baselines. We observe that across the three datasets, as the proportion of training sam-

[5]BERT-base-uncased: https://huggingface.co/bert-base-uncased;
RoBerta-base: https://huggingface.co/roberta-base.

Table 5: Experimental results of zero-shot scenario in terms of Accuracy (Acc) and Macro-F1 (F1). The best results are represented in **bold**. FT and PT respectively denote fine-tuning and prefix-tuning versions.

Datasets	ISEAR		TEC		CARER	
Metric	Acc	F1	Acc	F1	Acc	F1
BERT_{FT}	23.59	21.28	27.73	19.65	22.86	14.35
$\text{RoBERTa}_{\text{FT}}$	22.94	20.10	28.19	20.82	**24.06**	**15.39**
BERT_{PT}	23.60	21.21	30.03	23.84	22.35	14.12
$\text{RoBERTa}_{\text{PT}}$	30.77	29.43	29.28	21.97	23.11	14.05
VISPOR	**32.89**	**31.05**	**31.60**	**24.54**	23.96	15.16

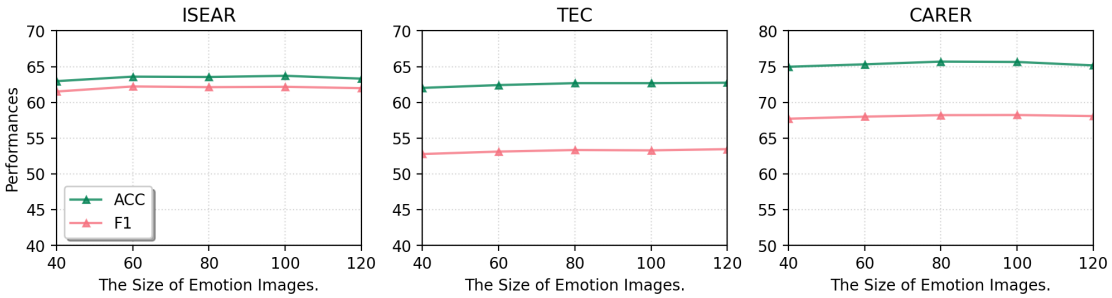

Figure 3: Performances of the VISPOR under different scales of the emotion image set.

ples increases, the VISPOR consistently surpasses the compared methods in terms of F1 score. This suggests that even under limited resources, utilizing cross-modal information of emotions through prefix-tuning can still improve task performance of CED.

4.5 Zero-shot Scenario.

As a number of emotion detectors are not based on the knowledge-rich PLMs, their performance in zero-shot scenarios without further training is quite poor. Therefore, we selected several PLM-induced methods including for comparison. The results are shown in Table 5. By analyzing the results, we observe that the VISPOR achieves the best performances in terms of Acc and F1 over the ISEAR and TEC datasets. This suggests that even without further training, the categorical emotion embeddings derived from the text-visual shared semantic space can significantly improve the ability of the VISPOR to understand text emotions.

4.6 Parameter Analysis.

Here, we analyze the impact of the number of selected emotion images to the VISPOR. Specifically, we respectively set the value of M_p to 40, 60, 80, 100, and 120. The performance in terms of Acc and F1-score under these different values is shown in Fig.3. We find that as the M_p increases, the VISPOR performs relatively stable over the ISEAR and CARER datasets. For the TEC dataset, the performance of the VISPOR shows a slight improvement. This indicates that introducing cross-modal visual information can enhance task performance. Moreover, it demonstrates the robustness of the VISPOR to visual information, as its performance is not significantly affected even when the number of emotional images is reduced.

5 Conclusion

In this paper, we present the VISPOR, which revisits the CED task from the cross-modal learning perspective. Considering the potential modality gap caused by directly encoding emotion images, we instead design abstract textual descriptions of emotions and align these class-wise texts with redundant emotion images, constructing a shared semantic space. By doing this, the VISPOR can seamlessly exploit such visual information of emotions by ingesting embeddings of these textual descriptions as prefix, improving the task performance of CED in the prefix-tuning paradigm. We evaluate the VISPOR over three CED benchmarks, and the results show that our method is significantly superior to the baselines. In the future, we will investigate how to effectively exploit the acoustic information of human beings, simultaneously improving the emotion understanding in both textual and visual modalities.

6 Limitation

In this paper, the proposed VISPOR primarily focuses on the scenario of Categorical Emotion States (CES). However, there is another emotion model, namely the Dimensional Emotion Space (DES), which represents various human emotions within a continuous space. In our future work, we will extend the VISPOR to the DES, aiming to enhance the model's generalization capability while also linking CES with DES within a deep semantic space.

On the other hand, the VISPOR coarse-grained utilizes emotion semantics within the visual space. However, when perceiving emotions in images, multiple factors are simultaneously considered, such as image brightness, color, and the spatial relationships between objects. Therefore, in future work, we will explore how to align such fine-grained image details with textual concepts in a shared semantic space, aiming to achieve better multimodal emotion representations.

References

Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. 2011. Mining social emotions from affective text. *IEEE transactions on knowledge and data engineering*, 24(9):1658–1670.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM Web Conference 2024*, page 1318–1326.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Erik Cambria, Paolo Gastaldo, Federica Bisio, and Rodolfo Zunino. 2015. An elm-based model for affective analogical reasoning. *Neurocomputing*, 149:443–455.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 29.

Lu Dai, Bang Wang, Wei Xiang, Minghua Xu, and Han Xu. 2022. A hybrid semantic-topic co-encoding network for social emotion classification. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 587–598. Springer.

Jiawen Deng and Fuji Ren. 2023. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67.

Eleanor Jack Gibson. 1969. Principles of perceptual learning and development.

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Lin Gui, Jia Leng, Gabriele Pergola, Yu Zhou, Ruifeng Xu, and Yulan He. 2019. Neural topic model with reinforcement learning. 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 3478–3483.

Jiale Han, Shuai Zhao, Bo Cheng, Shengkun Ma, and Wei Lu. 2022. Generative prompt tuning for relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3170–3185.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240.

Phil Katz, Matt Singleton, and Richard Wicentowski. 2007. Swat-mp: the semeval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 308–313.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2017. Emotic: Emotions in context dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 61–69.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, and Liu Wenyin. 2014. Towards building a social emotion detection system for online news. *Future Generation Computer Systems*, 37:438–448.

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.

Xianming Li, Zongxi Li, Haoran Xie, and Qing Li. 2021a. Merging statistical feature via adaptive gate for improved text classification. 35(15):13288–13296.

Zongxi Li, Xinhong Chen, Haoran Xie, Qing Li, Xiaohui Tao, and Gary Cheng. 2021b. Emochannel-sa: exploring emotional dependency towards classification task with self-attention mechanism. *World Wide Web*, 24:2049–2070.

Zongxi Li, Haoran Xie, Gary Cheng, and Qing Li. 2021c. Word-level emotion distribution with two schemas for short text emotion classification. *Knowledge-Based Systems*, 227:107163.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. 2023. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337.

Pingsheng Liu, Zhengjie Huang, Xiechi Zhang, Linlin Wang, Gerard de Melo, Xin Lin, Liang Pang, and Liang He. 2023. A disentangled-attention based framework with persona-aware prompt learning for dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13255–13263.

Xiao Liu, He-Yan Huang, Ge Shi, and Bo Wang. 2022. Dynamic prefix-tuning for generative template-based event extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5216–5228.

Andrew N Meltzoff and Richard W Borton. 1979. Intermodal matching by human neonates. *Nature*, 282(5737):403–404.

Saif Mohammad. 2012. # emotional tweets. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Chang Wang and Bang Wang. 2020. An end-to-end topic-enhanced self-attention network for social emotion classification. In *Proceedings of the web conference 2020*, pages 2210–2219.

Chang Wang, Bang Wang, Wei Xiang, and Minghua Xu. 2019. Encoding syntactic dependency and topical information for social emotion classification. In *Proceedings of the 42nd International ACM SIGIR Conference on research and development in information retrieval*, pages 881–884.

Zhihao Wen and Yuan Fang. 2024. Prompt tuning on graph-augmented low-resource text classification. *IEEE Transactions on Knowledge and Data Engineering*.

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

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Rui Hou, Yuxiao Dong, VG Vinod Vydiswaran, and Hao Ma. 2022. Idpg: An instance-dependent prompt generation method. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5507–5521.

Cheng-Ta Yang and Yi-Ling Chen. 2020. Dacnn: Dynamic weighted attention with multi-channel convolutional neural network for emotion recognition. In *2020 21st IEEE international conference on mobile data management (MDM)*, pages 316–321. IEEE.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394.

Jufeng Yang, Ming Sun, and Xiaoxiao Sun. 2017. Learning visual sentiment distributions via augmented conditional probability neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.