CLIP-SMP: sentiment recognition using CLIP with Synchronous Multimodal Prompts

Tianyang Shen
Department of computer science
Xi'an Jiaotong University
Xi'an, China
dazzyshen@126.com

***Abstract*—Sentiment recognition has always been a potential but challenging task in field of artificial intelligence. Especially with the zoom of the Internet of Thing(IoT) in recent years, the demand for high-resolution sentiment recognition is greatly growing. However, the absence of massive sentiment recognition datasets significantly obstacles its development. Recently Contrastive Language-Image Pre-training (CLIP) manifests us a new vision of compensating the lack of sentiment recognition datasets with the massive general knowledge contained in CLIP. But existing sentiment recognition models based on CLIP often only provide unimodal prompt individually or asynchronous prompts to CLIP, which might disrupt the balance of the multimodal structure in CLIP, finally impeding higher precision at sentiment recognition. In this paper, I propose CLIP-SMP, a sentiment recognition model using CLIP with lightweight synchronous multimodal prompts. Via experiments on two sentiment-recognition benchmarks, I prove the effectiveness and efficiency of CLIP-SMP, needing only 2.5M trainable parameters but reaching state-of-art.**

Keywords—sentiment recognition, CLIP, prompt, computer vision

I. INTRODUCTION

Sentiment recognition, as a pivotal task in artificial intelligence (AI), aims to decode human sentimental states from multimodal data such as text and image. In recent years, it has gathered significant attention due to its broad applications in human-computer interaction, mental health monitoring and other IoT-driven smart environments. Despite its potential, achieving steady and high-resolution sentiment recognition remains a challenge. Traditional AI approaches often rely on massive datasets and handcrafted features [7, 11], however paradoxically the absence of datasets in field of sentiment recognition has bogged down the its progress. The rapid proliferation of the Internet of Things (IoT) further amplifies this challenge. With billions of interconnected devices generating heterogeneous data streams—ranging from smart home sensors to wearable health monitors—there is an urgent demand for adaptive and accurate sentiment analysis systems capable of processing high-dimensional, real-time data while preserving contextual semantic information.

Recent advancements in large pre-training models, particularly Contrastive Language-Image Pre-training (CLIP), offer a promising pathway to mitigate these limitations [7]. CLIP's balanced architecture, which aligns visual and textual representations through contrastive learning on 400 million image-text pairs, embeds rich cross-modal knowledge that transcends domain-specific boundaries. This makes it a compelling candidate for sentiment recognition, where contextual alignment between modalities is effective. However, existing CLIP-based sentiment recognition frameworks often adopt unbalanced adaptation strategies. Some approaches naively fine-tune CLIP's entire architecture, disregarding its carefully pretrained architecture, compromising the robustness and generalization of CLIP [1, 2]. While some deploy unimodal prompts (text-only or image-only templates) or asynchronous prompts that fail to exploit and even disrupt CLIP's inherent cross-modal synergy [13, 14].

To address these limitations, this paper introduces CLIP-SMP (CLIP with Synchronous Multimodal Prompts), a novel framework that preserves CLIP's multimodal harmony while enabling efficient adaptation for sentiment recognition. Machine-learned prompts with frozen domain structure can avoid manually disruption and keep the robustness, generalization and neural semantic information learned at training of large datasets [16, 17, 18]. At its core, CLIP-SMP employs lightweight, learnable prompt pairs *Prompt-Wv* and *Prompt-Wt* transferred from a common initial prompt *Prompt-initial* that operate synchronously across visual and textual modalities. These prompts act as "steering vectors", collaboratively guiding CLIP's frozen backbone to refine its attention toward sentiment-related features shared by image and text both, without distorting its pretrained knowledge. The synchronization mechanism ensures that textual prompts and visual prompts evolve in tandem during training, maintaining cross-modal sentimental coherence and help the model to distribute more attention on information shared by tow modalities both, especially sentimental information. Remarkably, this approach requires only 2.5M trainable parameters—a mere 0.3% of CLIP's total parameters—while MaPLe, another multimodal prompting CLIP-based model needs 19.3M trainable parameters [14]. Moreover, in the experiments CLIP-SMP outperforms state-of-the-art prompting model Emotion-CLIP and CLIP, also reaching state-of-the-art [5, 7]. This makes it exceptionally resource-efficient for deployment on IoT edge devices with limited computational budgets.

The efficacy of CLIP-SMP is rigorously validated through experiments on two benchmark datasets: Emotic [19] and MELD [20]. Results demonstrate that CLIP-SMP not only reaches state-of-the-art but also exhibits great robustness in cross-dataset evaluations. These achievements highlight a broader paradigm: instead of overhauling foundation models for downstream tasks, strategically guiding their multimodal alignment through lightweight, coordinated interventions can yield disproportionate improvements in efficiency and accuracy.

This work advances the field in three key dimensions:

(1) It identifies and rectifies the modality asynchrony prompting problem in CLIP-based sentiment recognition.

(2) It establishes a resource-efficient framework for adapting large vision-language models to data-scarce scenarios.

(3) It provides empirical evidence that balanced multimodal synchronization—not just scale—dictates the success of foundation models in affective computing.

For IoT ecosystems increasingly reliant on ambient emotional intelligence, CLIP-SMP offers a scalable blueprint to harness general-purpose big-scale pre-trained AI models without compromising their inherent strengths.

II. Related word

A. Contrastive Language-Image Pre-training

Contrastive Language-Image Pre-training(CLIP) [7] is a groundbreaking multimodal model that learns visual representations through natural language supervision. Its core idea is to train a system to associate images with their corresponding textual descriptions by leveraging a contrastive learning objective. Instead of traditionally relying on manually annotated categorical labels, CLIP uses 400 million image-text pairs collected from the internet, enabling it to learn a broad spectrum of visual concepts directly from raw text.

CLIP jointly trains two encoders, an image encoder (ResNet [12] or Vision Transformer [10]) and a text encoder (Transformer [21]), to maximize the cosine similarity of embeddings for matched image-text pairs within a batch while minimizing similarity for mismatched pairs. This contrastive objective, inspired by InfoNCE loss [22], allows the model to learn a shared embedding space where semantically related images and texts are aligned. CLIP has some main advantages as followed

(1) Zero-shot Transfer: CLIP can generalize to unseen tasks without dataset-specific fine-tuning. For example, it matches ResNet-50's accuracy on ImageNet (76.2%) without any extra training.

(2) Multitask Capability: It excels in diverse tasks like OCR, action recognition, and geo-localization by dynamically synthesizing classifiers via text prompts.

(3) Robustness: CLIP shows stronger performance under natural distribution shifts compared to traditional supervised models, as it avoids overfitting to narrow label sets.

All advantages showed above indicate that CLIP is a bridge between vision and language, offering a flexible framework for task-agnostic visual understanding while highlighting the potential of scalable, language-driven supervision in AI. Let me foresee the great potential to transfer CLIP to sentiment recognition task.

B. Directly using CLIP for zero-shot sentiment recognition

Due to the rich image-text alignment captured in the CLIP, directly applying CLIP for zero-shot sentiment recognition yields strong performance. Xin et al. [13] argue that fine-tuning might compromise CLIP's structural integrity and degrade its effectiveness, leading many researchers to adopt zero-shot approaches. Bustos et al. [1] found that zero-shot CLIP achieves near state-of-the-art (SOTA) performance in sentiment recognition and even surpasses few-shot CLIP and other SOTA models on large-scale datasets. This success might stem from CLIP's massive natural language supervision during pretraining, which endows it with exceptional robustness and generalization. Such an approach fully leverages CLIP's inherent knowledge while preserving its pretrained architecture.

However, zero-shot CLIP underperforms compared to few-shot or fine-tuned variants when tackling specialized, complex, or abstract tasks or small-scale datasets [1, 2, 7]. This limitation arises because while CLIP's pretrained knowledge is broad and generalizable, but lacks domain-specific expertise tailored to niche applications.

C. Enhanced pretraining based on CLIP

To improve performance in sentiment recognition—a task requiring abstraction and specialization — researchers worldwide have proposed sentiment recognition models built upon further pretraining of CLIP. Main research directions include:

(1) Integrating CLIP with other pretrained models

Devillers et al. [8] note that CLIP, which focuses on extracting cross-modal (image-text) shared information, may lose unique unimodal features critical for single-modality tasks, leading to suboptimal performance compared to specialized unimodal models. Bielawski et al. [9] further observe that combining vision-only and text-only large models can outperform CLIP in certain domains, as their combined knowledge may better cover both shared and modality-specific information. To address this, researchers propose integrating CLIP with unimodal models.

For example, Zichao Nie et al. [3] fuse features from ViT [10] for images, RoBERTa [11] for text, and CLIP for cross-modal alignment to predict sentiments. Lu et al. [4] first extract image and text features separately using ResNet and RoBERTa, then refine and fuse them via CLIP.

(2) Efficient contrastive learning strategies

While CLIP's core strength lies in contrastive learning between images and text using natural language supervision, Bielawski et al. [9] and Devillers et al. [8] argue that language supervision does not always outperform traditional supervised methods. However, Bielawski et al. [9] find that for human-centric tasks such as sentiment recognition, CLIP surpasses unimodal models like ViT or RoBERTa, as human-centric semantics are better captured in cross-modal shared information. This shows us the bright foreground of further pretraining CLIP-based models specialized for sentiment recognition task.

To help CLIP capture more human-centric especially sentiment semantics, researchers optimize CLIP's contrastive learning to focus on affective cues. Zhang et al. [5] leverage a pretrained language model to predict sentiments from dialogue text first, then adjust contrastive weights between video-text pairs based on predicted sentiment labels, guiding the model to prioritize sentiment-related semantics during training.

(3) Guiding contrastive learning with prompts

Radford et al. [7] demonstrate that text prompts such as "A photo of a {label}" significantly improve classification by reducing semantic ambiguity. Extending this idea, researchers propose using image prompts alongside text prompts to steer contrastive learning toward sentiment-specific features.

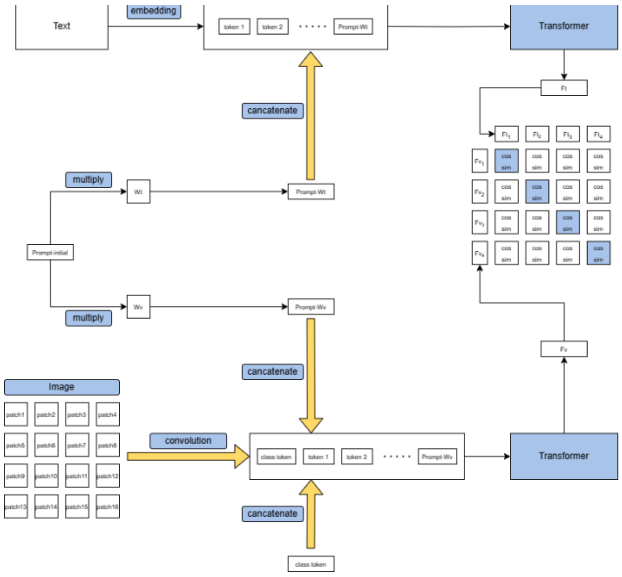

Fig. 1. The structure of CLIP-SMP

Deng et al. [6] reuse CLIP's text encoder to encode multiple prompt variations and sentiment synonyms, selecting the most image-aligned prompts to maximize the shared sentiment semantics from contrastive learning. Zhang et al. [5] incorporate positional features as image prompts to strengthen subject-text alignment. Xin et al. [13] and Khattak [14] emphasize that multimodal prompting preserves CLIP's symmetric contrastive structure better than unimodal prompting, making it a promising avenue for future work, which is the very direction I am researching.

These approaches aim to refine CLIP's pretrained knowledge for sentiment recognition while maintaining its generalization power, balancing broad pretraining with task-specific adaptation.

D. *Contrastive learning from modalities beyond images and text*

CLIP's unprecedented capabilities stem from its contrastive learning between images and text. To further enhance its robustness and generalization, researchers have explored integrating additional modalities, such as audio, into the contrastive learning framework.

For instance, Guzhov et al. [15] pretrained an audio feature extractor and performed contrastive learning between audio, image, and text modalities. While this approach improved the quality of audio feature extraction, it had limited impact on enhancing image or text feature extractors. Nonetheless, their work highlights the feasibility of incorporating audio modality for emotion recognition tasks, broadening CLIP's applicability to multimodal scenarios.

III. Method

My core idea is offering CLIP lightweight prompts to guide CLIP's contrastive learning focus on sentiment semantics shared by image and text both. Meanwhile, to keep the intact and balanced structure contained affiliation between image and text, I keep CLIP itself frozen and train the prompts individually. The whole structure of CLIP-SMP is shown in Fig.1.

A. *Image encoder*

The image encoder used in CLIP-SMP is Vision Transformer initialized with parameters of CLIP. Vision Transformer at first divides image *X* into fixed-size, non-overlapping patches(16*16). Each patch is flattened into a vector through convolution. Then these patch vectors are linearly projected into lower-dimensional embeddings. A learnable "[class]" token is prepended to the sequence to aggregate global information for classification. Also, positional embeddings are added to the patch embeddings to retain spatial relationships, as Transformers lack inherent spatial awareness. Now we get a sequence of embeddings *Zv* = [z_{class}, z_1, z_2, …] which is subsequently passed through Transformer encoder to get ultimate image feature *Fv*.

B. *Text encoder*

The text encoder used in CLIP-SMP is a Transformer. Word sequence *T* is embedded at first into a token sequence *Zt*. The width of *Zt* is not equal the width of features, so the *Zt* is projected to the width of features next. Finally pass the projected *Zt* through Transformer to get ultimate text feature *Ft*.

C. *Prompting*

Image encoder and text encoder described above are original encoders used in CLIP's original paper. CLIP-SMP's progress of encoding is a little different because of additional prompts.

Initially, a parameter *Prompt-initial(1x512)* is generated randomly. Matrixes *Weight-Wv(512x768)* and *Weight-Wt(512x512)* are generated at same time. Then multiply *Weight-Wv* and *Weight-Wt* individually with *Prompt-initial* to produce *Prompt-Wv(1x768)* and *Prompt-Wt(1x512)*.

$$Prompt\text{-}Wv = Prompt\text{-}initial \ @ \ Weight\text{-}Wv \quad (1)$$

$$Prompt\text{-}Wt = Prompt\text{-}initial \ @ \ Weight\text{-}Wt \quad (2)$$

To implement the produced prompts, *Prompt-Wv* and *Zv* and concatenated into [z_{class}, z_1, z_2, …, *Prompt-Wv*]. *Prompt-Wt* and *Zt* are concatenated into [z_{class}, z_1, z_2, …, *EOS*, *Prompt-Wt*]. Concatenated *Zv* and *Zt* are really used in the stream.

D. *Contrastive Learning*

After extracting features *Fv* and *Ft*, to contrastive learn the affiliation between text and image, I use the loss function from CLIP's original paper, a simple but effective cosine similarity loss.

$$L_v = \frac{1}{N}\sum_{i=1}^{N}\left[-log\frac{\exp(\langle Fv_i, Ft_i\rangle)}{\sum_{j=1}^{N}\exp(\langle Fv_i, Ft_j\rangle)}\right] \quad (3)$$

$$L_t = \frac{1}{N}\sum_{i=1}^{N}\left[-log\frac{\exp(\langle Fv_i, Ft_i\rangle)}{\sum_{j=1}^{N}\exp(\langle Fv_j, Ft_i\rangle)}\right] \quad (4)$$

$$loss = \frac{L_v + L_t}{2} \quad (5)$$

In loss function, $\langle Fv_i, Ft_i\rangle$ means cosine similarity between Fv_i and Ft_i. The loss function optimizes contrastive alignment by pulling each image's embedding toward its paired text while pushing it away from embeddings of all other samples within the batch.

IV. Experiment

I train CLIP-SMP on two sentiment recognition benchmark, Emotic and MELD. I compare performances of different training strategies: only training on Emotic, only

training on MELD and training on two datasets both. I find some unexpected results.

A. Datasets

(1) Emotic

Emotic [19] is an image dataset for sentiment recognition in context, comprising 23,571 images of 34,320 annotated individuals in real world. Each sample is annotated with 26 discrete categories.

(2) MELD

MELD [20] is an extension to the EmotionLines [23], which is a sentiment corpus of conversations initially proposed in the field of NLP. MELD offers the same dialogue examples in EmotionLines and includes audio and visual modalities along with the text. It contains around 1,400 dialogues and 13,000 utterances from the *Friends* tv show, where each sample is annotated with 7 discrete categories.

B. Linear Probe

I use linear probe to evaluate the performance of CLIP-SMP and compared it with other models. In linear probing,

TABLE I. LINEAR PROBE RESULTS

Model	Dataset for test	
	Emotic	***MELD***
CLIP	32.240	43.908
CLIP-SMP trained on Emotic	33.414	43.563
CLIP-SMP trained on MELD	34.109	45.977
CLIP-SMP trained on the both	33.919	44.100
EmotionCLIP	32.91	48.28

[a.] Evaluation metric is accuracy

CLIP-SMP is frozen, and a simple linear layer is added on top of the model's frozen features. Only this layer is trained on labeled data for downstream task (sentiment recognition).

By freezing the model, it isolates the evaluation to the quality of pretrained features, not the model's ability to adapt to new data. Many models use linear probe as methods to evaluate their models. So, such method is not only cheap and fast, but also easy to compare with other models.

C. Result

All CLIP-SMPs are trained for 5 epoch with weight decay 0.2. Linear probing results are shown in TABLE I.

Compared with CLIP, CLIP-SMPs whatever the dataset used for training is outperform the performance of CLIP, which proves the effectiveness of the lightweight prompting. Compared with EmotionCLIP, a state-of-the-art sentiment recognition model, CLIP-SMP outperforms on Emotic, but fails on MELD. This might be caused by the lack of datasets, restricted by limiting time. What's interesting is, that the highest accuracy on Emotic is obtained by CLIP-SMP only trained on MELD. This definitely proves the robustness of CLIP-SMP and the correctness freezing rather than fine-tuning CLIP. Also, through this phenomenon, we can be convinced that sentimental semantics are truly extracted into prompts and CLIP's attention is truly focused on sentiment via synchronous multimodal prompts.

V. CONCLUSION

In this paper, we introduced CLIP-SMP, a novel framework that leverages CLIP's pretrained multimodal capabilities through lightweight synchronous multimodal prompts for sentiment recognition. By designing learnable prompt pairs that operate in synchronously across visual and textual modalities, CLIP-SMP effectively guides CLIP's frozen backbone to focus on sentimental semantics while preserving its intrinsic cross-modal alignment. This approach addresses the critical limitations of existing methods, such as modality asynchrony and structural disruption, by ensuring balanced and coordinated adaptation.

Experimental results on the Emotic and MELD benchmarks demonstrate that CLIP-SMP achieves state-of-the-art performance with remarkable efficiency, requiring only 2.5M trainable parameters (0.3% of CLIP's total parameters). This efficiency makes it particularly suitable for deployment in resource-constrained IoT environments, where computational budgets are limited but demand for real-time, high-resolution sentiment analysis is growing. Furthermore, cross-dataset evaluations highlight the framework's robustness and generalization ability, underscoring the importance of maintaining CLIP's pretrained knowledge during adaptation.

This work advances the field of affective computing by establishing a paradigm where strategic, lightweight interventions rather than exhaustive retraining can unlock the full potential of foundation models like CLIP. Future research could explore extending CLIP-SMP to incorporate additional modalities, refining prompt synchronization mechanisms, or adapting the framework to other human-centric tasks such as intention recognition or mental health monitoring. By bridging the gap between general-purpose pretraining and domain-specific needs, CLIP-SMP paves the way for scalable, efficient, and context-aware emotional intelligence in next-generation AI systems.

REFERENCES

[1] Cristina Bustos, Carles Civit, et al. On the use of Vision-Language models for Visual Sentiment Analysis: a study on CLIP, 11th International Conference on Affective Computing and Intelligent Interaction (ACII), 2023

[2] Alessandro Bondielli, Lucia C. Passaro. Leveraging CLIP for Image Emotion Recognition, Ceur workshop proceedings, 2021

[3] Zichao Nie. Feature-Attentive Multimodal Emotion Analyzer with CLIP, International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), 2023

[4] Xintao Lu, Yonglong Ni, Zuohua Ding. Cross-Modal Sentiment Analysis Based on CLIP Image-Text Attention Interaction, International Journal of Advanced Computer Science & Applications, 2024

[5] Sitao Zhang, Yimu Pan, James Z. Wang. Learning Emotion Representations from Verbal and Nonverbal Communication, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023

[6] Sinuo Deng, Lifang Wu, et al. Learning to compose diversified prompts for image emotion classification, Computational Visual Media, 2024

[7] Alec Radford, Jong Wook Kim, et al. Learning Transferable Visual Models From Natural Language Supervision, International conference on machine learning(ICML), 2021

[8] Benjamin Devillers, Bhavin Choksi, et al. Does language help generalization in vision models?", Proceedings of the 25th Conference on Computational Natural Language Learning, 2021

[9] Romain Bielawski, Benjamin Devillers, et al. When does CLIP generalize better than unimodal models? When judging human-centric

concepts", Proceedings of the 7th Workshop on Representation Learning for NLP, 2022

[10] Alexey Dosovitskiy, Lucas Beyer, et al, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, International Conference on Learning Representations, 2020

[11] Yinhan Liu, Myle Ott, et al, RoBERTa: A Robustly Optimized BERT Pretraining Approach, International Conference on Learning Representations, 2020

[12] Kaiming He, Xiangyu Zhang, et al. Deep Residual Learning for Image Recognition, Computer Vision and Pattern Recognition, 2015

[13] Yi Xin, Junlong Du et al. MmAP:Multi-modal Alignment Prompt for Cross-domain Multi-task Learning, Proceedings of the AAAI Conference on Artificial Intelligence, 2024

[14] Muhammad Uzair Khattak, Hanoona Rasheed, et al. MaPLe: Multi-modal Prompt Learning, Conference on Computer Vision and Pattern Recognition (CVPR), 2023

[15] Andrey Guzhov, Federico Raue et al. AudioCLIP: Extending CLIP to Image, Text and Audio, International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022

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

[17] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5206–5215, 2022.

[18] Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Gold stein Tom, Anandkumar Anima, and Xiao Chaowei. Test time prompt tuning for zero-shot generalization in vision language models. 2022.

[19] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In CVPR, July 2017.

[20] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meet ing of the Association for Computational Linguistics, pages 527–536, Florence, Italy, July 2019. Association for Com putational Linguistics.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. NeurIPS, 30, 2017.

[22] Oord, A. v. d., Li, Y., and Vinyals, O. Representation learn ing with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018

[23] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun Wei Ku, et al. Emotionlines: An emotion corpus of multi party conversations. arXiv preprint arXiv:1802.08379, 2018.