Prompt Link Multimodal Fusion in Multimodal Sentiment Analysis

Kang Zhu[1], *Cunhang Fan*[1], *Jianhua Tao*[2,3], *Zhao Lv*[1]

[1]School of Computer Science and Technology, Anhui University, Hefei 230601, China,
[2]Department of Automation, Tsinghua University, China,
[3]Beijing National Research Center for Information Science and Technology,Tsinghua University

E22201061@stu.ahu.edu.cn, cunhang.fan@ahu.edu.cn, jhtao@tsinghua.edu.cn, kjlz@ahu.edu.cn

Abstract

Multimodal sentiment analysis aims to analyze sentiment by integrating information from various modalities. Combining different modalities can be challenging due to their inherent differences in distance. While researchers employ complex methods to reduce distances, connecting multiple modalities remains limited. In this paper, we introduce the technique of prompt learning and propose the Prompt Link Multimodal Fusion (PLMF), which consists of three components: Channel Prompt Link (CPL), Spatial Prompt Link (SPL), and Fusion Result Constraints (FRC). CPL facilitates fine-grained sentiment feature linkage in the channel dimension, while SPL connects overall sentiment semantic information in the temporal dimension. Due to the randomness of connecting vectors, FRC is proposed to constrain the linkage toward the direction of optimal fusion results. Through the collaborative efforts of these three modules, PLMF achieves state-of-the-art results on three publicly available datasets.

Index Terms: multimodal sentiment analysis, modal fusion, link multimodal

1. Introduction

Multimodal Sentiment Analysis (MSA) employs a combination of human sensory modalities, encompassing elements such as audio, images, and text, to facilitate the predictive modeling of human sentiment scores. Its applications are extensive, encompassing various domains such as health[1], human-computer interaction[2][3], marketing management[4][5], and social media analysis[6][7]. Unimodal approaches face significant challenges in recognizing nuanced sentiments such as irony. Hence, MSA possesses advantages that are unparalleled by unimodal methods. To further enhance the accuracy of sentiment analysis and the utilization of multimodal information, multimodal fusion plays a crucial role. However, the obstacle of fusion arising from intermodal distance differences presents a formidable challenge in the field of sentiment analysis[8].

In MSA, methods for addressing intermodal distance can be broadly categorized into two classes: attention-based and contrastive learning-based. Attention-based approaches focus on extracting relevant information while disregarding intermodal distances[9][10]. While attention-based methods offer the advantage of disregarding intermodal distances, challenges arise when modal features are at a lower level, as they often contain redundant information and noise. At higher levels of modal features, the use of attention mechanisms faces difficulties in capturing a comprehensive set of features due to the potential loss of low-level private information within each modality. This complexity hinders attention mechanisms from effectively focusing on pertinent information. On the other hand, contrastive learning-based methods aim to design approaches that minimize the distances between modalities [11][12][13]. Researchers leverage the characteristics of contrastive learning to design positive and negative sample pairs within modalities, between modalities, and across samples. This approach aims to minimize intermodal distances from various perspectives. Despite the promising results achieved by contrastive learning-based methods, the reduction of distances may not reach the theoretical limit of "0", and achieving optimal multimodal fusion remains a challenging task. Reducing intermodal distances and extracting cross-modal features alone may not lead to optimal connection relationships between different modalities. The prompt learning approach of large-scale language models presents a potential solution to achieve optimal connection relationships.

Prompt learning has demonstrated significant success in natural language processing[14][15][16] and computer vision domains[17][18]. In prompt learning, the prompt component takes various forms, broadly categorized into two types: hard prompt and soft prompt. Hard prompt involve incorporating fixed words or cropped images into the input, establishing a connection between the input and the model's knowledge. Given the various drawbacks associated with hard prompts, the concept of soft prompt is proposed. Soft prompt entail the addition of continuous learnable vectors, enabling automatic fitting to connect the input and the model, making them more versatile and flexible. The prompt component assists in achieving desirable output results for inputs beyond the model's perceptual range, facilitating the connection of inputs outside the model's perception range. Continuous learnable prompts enable automatic learning to address out-of-domain distances. In multimodal sentiment analysis, where each modality represents a distinct domain, similar out-of-domain distances exist between them, hindering modality fusion. We attempt to mitigate intermodal distances and facilitate multimodal fusion through the application of prompt learning techniques in our research.

In this paper, inspired by prompt learning techniques, we propose the Prompt Link Multimodal Fusion (PLMF) framework. It is designed to automatically learn and address intermodal distance issues, thereby facilitating fusion between modalities. The framework consists of three components: Channel Prompt Link (CPL), Spatial Prompt Link (SPL), and Fusion Result Constraints (FRC). CPL facilitates fine-grained sentiment feature connections in the channel dimension, enabling a more comprehensive acquisition of sentiment information. SPL focuses on connecting overall semantic information across modalities in the temporal dimension. The combined action of CPL and SPL achieves both localized and globalized comprehensive connections. To further constrain the learning direction of the automatically learnable vectors, we propose the

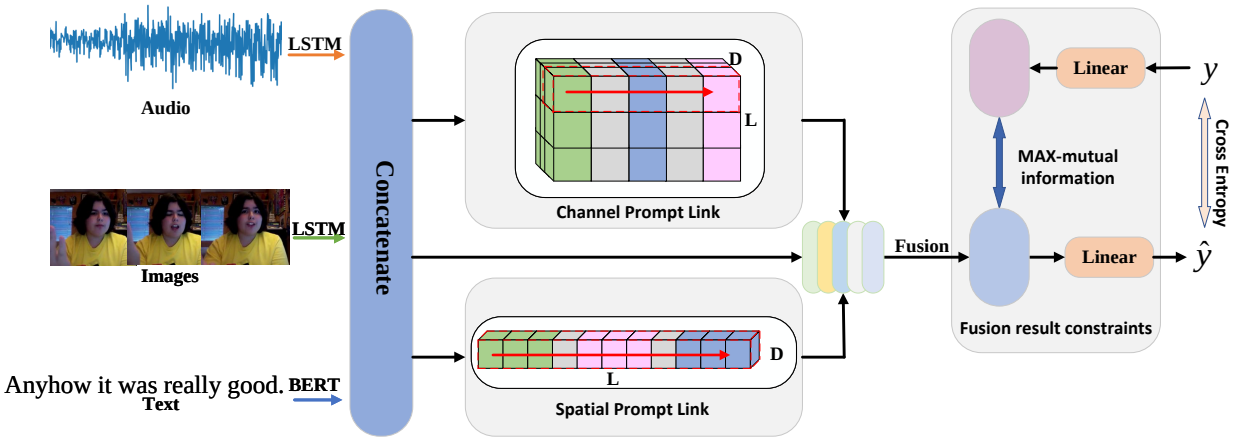


Figure 1: *The overall architecture of the PLMF model. PLMF mainly consists of CPL,SPL,FRC. $\hat{y}$ represents the predicted label, and y represents the true label.*

FRC module. By elevating the dimensions of the true labels, FRC optimizes the fusion results and imposes constraints on the fusion results generated by the front end, directing the learnable vectors toward the optimal fusion result direction. Through experimentation, it has been observed that soft prompt technology effectively addresses intermodal distance issues, and concurrently, the model achieves optimal experimental results.

2. Method

2.1. Overall Architecture

The comprehensive framework of PLMF is depicted in Fig 1. Concerning modal feature extraction, the text (t) modality employs the pre-trained BERT model [19] for feature extraction from raw textual data. Meanwhile, the audio (a) and visual (v) modalities leverage Long Short-Term Memory (LSTM) architecture [20] to extract features. h_t is the head embedding extracted from the output of the last layer of BERT; h_m is the feature of the last time step of LSTM, $m \in \{a, v\}$. They serve as inputs to the model. PLMF consists of three major modules: CPL, SPL, and FRC. CPL and SPL are responsible for establishing connections between different modalities, while FRC guides modality connections. The model predicts results by integrating intermediate representations of audio, images, and text, along with the outputs from CPL and SPL.

2.2. Channel Prompt Link

The CPL module is responsible for connecting fine-grained emotional features by linking different modality-specific sentimental information features. This facilitates a more comprehensive integration of sentimental information, aiming to minimize the loss of information during the extraction of cross-modal information using attention mechanisms. As illustrated in Fig 1, it employs randomly initialized learnable prompt vectors P for automatic adaptive connections across different modality channels. The individual feature vectors undergo channel-wise elevation and stacking, followed by the utilization of a convolutional neural network for local fine-grained emotional feature connections across channels. The dimension is elevated to four dimensions, akin to a picture being divided into three segments, with the learnable vectors serving as a cohesive agent across the gaps between them.

$$X_m = Conv1d(h_m) \tag{1}$$

$$H_{CLP} = Conv2d(stack((X_t, P, X_a, P, X_v), dim = 1)) \tag{2}$$

Where X_m represents the result after channel-wise elevation, P is the prompt vector, $m \in \{t, a, v\}$.

2.3. Spatial Prompt Link

SPL is responsible for connecting overall sentimental information across modalities, and it operates in the temporal dimension. Despite different modalities expressing the same semantic sentiment, there can be significant differences in the form of sentimental semantics between modalities. SPL is designed to reduce this distance. As illustrated in Figure 1, SPL facilitates temporal connections and convolution operations.

$$S = Cat([t, P, a, P, v], \text{dim} = 1) \tag{3}$$

$$H_{SPL} = Relu(Conv1d(S)) \tag{4}$$

2.4. Fusion result constraints

As the learnable prompts are vectors learned automatically, their learning direction is undetermined after random initialization. To find the optimal learning direction, we employ FRC to enforce reverse label constraints. Utilizing contrastive learning[21][22], we maximize the mutual information between the model's fusion results and the reverse features of the labels, guiding the learnable vectors towards optimal fusion results.

$$F = Relu(Linear(y)) \tag{5}$$

Where y represents the ground truth label.
Given the significant success of [21] contrastive learning design, we introduce its loss into our framework

$$CL\_Loss = -log\frac{\exp(q \cdot k^+/\tau)}{\sum_{i=1}^{b} \exp(q \cdot q^i/\tau)} \tag{6}$$

where b is batch size; the query sample q and positive key sample k^+,τ is the temperature coefficient;

The fusion is performed on intermediate results from various sources to maximize mutual information.

$$H = Cat([t, a, v, H_{CPL}, H_{SPL}], \dim = 1) \tag{7}$$

$$\hat{F} = Linear(H) \tag{8}$$

$$MAX\_MILoss = CL\_Loss(\hat{F}, F) \tag{9}$$

2.5. Sentiment Intensity Prediction

With the assistance of each module, the final fusion results are predicted, and cross-entropy loss is computed.

$$\text{CrossEntropy\_Loss: } -\sum_{i=1}^{K} y_i \cdot \log(\hat{y}_i) \tag{10}$$

Where K is the number of categories, y is the true label, and $\hat{y}$ is the predicted value.
The overall loss of PLMF is given by:

$$Loss = \lambda MAX\_MILoss + CrossEntropy\_Loss \tag{11}$$

where λ is a hyperparameter.

3. Experiments

3.1. datasets

PMLF utilizes popular publicly available datasets in the MSA domain, namely the CMU-MOSI[23] dataset, a key benchmark in MSA research comprising 93 YouTube monologues with 26,295 words in 2,199 opinion video utterances. Sentiment strength labels range from -3 to +3. CMU-MOSEI [24] is a large MSA dataset with 23,454 YouTube video clips covering 250 topics from 1,000 speakers. Utterances include movie review topics annotated with sentiment scores (-3 to +3) and 6 sentiment categories. CH-SIMS [25] is a Chinese MSA dataset with 2,281 video clips from movies, TV series, and variety shows. It includes unified multimodal and independent unimodal annotations, with sentiment scores ranging from -1 to 1. Their respective divisions are outlined in Table 1.

Table 1: *Dataset split.*

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856
CH-SIMS	1368	456	457	2281

3.2. Experiment settings

We fine-tuned the pre-trained models, bert-base-uncased and bert-base-chinese, with a learning rate of 5e-5. The parameters for CMU-MOSI, CMU-MOSEI, and CH-SIMS are as follows: model learning rates are 9e-4, 6e-6, 1e-4, λ values are 0.6, 0.05, 0.02, and batch sizes are 144, 448, and 128, respectively. The convolution kernel size for CPL is set to 5, SPL convolution kernel sizes are 5, 32, and 3, and FRC dimension elevation is 128. The dimensions for Bert and LSTM outputs are 768 and 64, respectively. We conducted training and experimental analysis on an RTX 3090, using evaluation metrics such as mean absolute error (MAE), Pearson correlation (Corr), binary classification accuracy (Acc-2), and F1 scores.

3.3. Comparison

Table 4 and Table 2 present the comparison results between PLMF and other classical models. From various metrics, it is evident that PLMF achieves optimal results in all indicators with a significant performance improvement. This further validates the effectiveness and feasibility of PLMF, indicating that the use of soft prompt to address intermodal distance issues is both effective and viable. Simultaneously achieving optimal results with very few parameters, we believe that soft prompt technology proves to be a promising approach to addressing intermodal distance issues in multimodal scenarios. PLMF not only outperforms the state-of-the-art results, but it also utilizes significantly fewer parameters compared to models using attention mechanisms. The total parameter count of CPL, SPL, and FRC combined in PLMF is only 0.56M.

Table 2: *Results on CH-SIMS.*

Model	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
TFN [26]	43.22	59.1	78.38	78.62
LMF [27]	44.12	57.59	77.77	77.88
MulT [28]	45.32	56.41	78.56	79.66
Self-MM [29]	42.50	**59.52**	80.04	80.44
PLMF(Ours)	**42.44**	58.58	**82.03**	**81.36**

3.4. Ablation Study

The results of the ablation experiments, as illustrated in Table 3, distinctly reveals a substantial degradation in the performance of PLMF. This underscores the indispensable nature of CPL, SPL, and FRC components. Their synergistic interplay is pivotal in fostering effective intermodal connections, enabling the model to acquire exemplary representations and, consequently, achieving optimal fusion outcomes. The results show a decrease of 1.8% and 2.3% without CPL or SPL, indicating that the overall semantics of different modalities are more crucial for classification compared to sentimental details. The 2.3% decrease without FRC suggests the crucial importance of the learning direction for soft prompts, and without constraints, modal connections cannot be effectively established.

Table 3: *Ablation study of PLMF on CMU-MOSI; Channel Prompt Link (CPL), Spatial Prompt Link (SPL) and Fusion result constraints (FRC).*

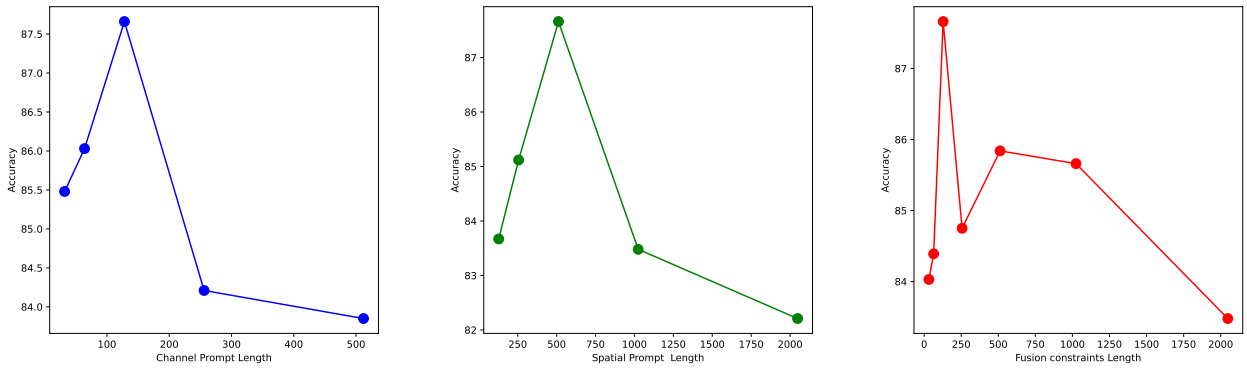
Description	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
PLMF	**0.704**	**0.807**	**85.4/87.6**	**85.2/87.6**
w/o CPL	0.732	0.804	84.0 / 85.8	83.9 / 85.8
w/o SPL	0.751	0.790	82.8 / 84.3	82.7 / 84.3
w/o FRC	0.739	0.802	82.8 / 84.3	82.8 / 84.4

3.5. Length Analysis

In our exploration of each module within the PLMF framework, as illustrated in Figure 3, a and b represent the exploration of prompt lengths in the CPL and SPL modules, respectively, while c indicates the impact of enhancing the real label dimension on the model. The trends observed in a and b are similar, suggesting that shorter prompt lengths result in inadequate connections between modalities, leading to suboptimal performance. As the lengths increase, the performance improves. However, excessively long lengths introduce too many random values during initialization, hindering the FRC's ability to provide effective constraints and leading to diminished

Table 4: *Results on CMU-MOSI and CMU-MOSEI; All models use bert-base-uncased as the text encoder; In Acc-2 and F1-Score, the left of the "/" is calculated as negative/non-negative and the right is calculated as negative/positive. Performance Comparison between DVMI and baselines on CMU-MOSI and CMU-MOSEI datasets. Baseline results are sourced from [30] and [31].*

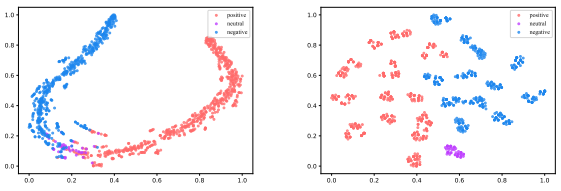
	CMU-MOSI				CMU-MOSEI			
Models	MAE ↓	Corr↑	Acc-2↑	F1↑	MAE ↓	Corr↑	Acc-2↑	F1↑
TFN [26]	0.925	0.662	78.3 / 80.2	78.2 / 80.1	0.570	0.716	81.0 / 82.6	81.1 / 82.3
LMF [27]	0.931	0.670	77.5 / 80.1	77.3 / 80.0	0.568	0.727	81.3 / 83.7	81.6 / 83.8
MulT [28]	0.918	0.685	79.0 / 80.5	79.0 / 80.5	0.564	0.732	81.3 / 84.0	81.6 / 83.9
MISA [32]	0.752	0.784	81.8 / 83.5	81.7 / 83.5	0.550	0.758	81.6 / 84.3	82.0 / 84.3
MMIM [33]	0.738	0.781	83.0 / 85.1	82.9 / 85.0	**0.547**	0.752	81.9 / 85.1	82.3 / 85.0
PLMF(Ours)	**0.704**	**0.807**	**85.4/87.6**	**85.2/87.6**	0.552	**0.773**	**83.8/86.8**	**84.1/86.7**



(a) (b) (c)

Figure 2: *a and b represent the lengths of the prompt parts for CPL and SPL, respectively, while c indicates the dimensionality increase for the true labels.*

performance. For the real label, which contains abundant information, a smaller dimension leads to crowded information distribution in a shorter vector, making mutual information constraints challenging. Conversely, an overly large dimension disperses information too widely, with some units lacking information, potentially misleading the model and causing errors in constraint direction. Therefore, a balanced and moderate approach to determining the lengths of prompt and real label dimension enhancement is advocated for optimal results.



(a) *model fusion results* (b) *real label dimension improve*

Figure 3: *t-SNE [34]visualization of multimodal fusion result in the embedding space on CMU-MOSI training set.*

3.6. Further Analysis

We further visualized the fusion results using t-SNE[34], where a represents the model's output, and b represents the results of enhancing the real label dimensions. The data distribution in b closely matches the distribution of sentiment data, indicating that appropriate label dimension enhancement does not lead to information loss. With the assistance of each module, the distribution results are excellent. It precisely mirrors the movement of the same-colored distribution in b towards both ends, demonstrating the effectiveness of FRC on CPL and SPL. This further confirms the rationality, effectiveness, and feasibility of PLMF.

4. Conclusion

In this paper, we introduce the prompt learning technique to address intermodal distance issues and propose the PLMF framework. We design Spatial Prompt Link (SPL) and Channel Prompt Link (CPL) to establish connections for fine-grained sentiment features and overall semantic information between modalities. Due to the automatic learning direction of the prompt, which is challenging to determine, we propose Fusion Result Constraints (FRC) constrain them to learn and connect multiple modalities in the direction of optimal fusion results. Experimental results validate the feasibility and effectiveness of our approach. The flexibility of prompt technology allows for various applications, and in the future, we will explore its additional uses.

5. Acknowledgements

This work is supported by the STI 2030—Major Projects (No. 2021ZD0201500), the National Natural Science Foundation of

China (NSFC) (No.62201002), Distinguished Youth Foundation of Anhui Scientific Committee (No. 2208085J05), Special Fund for Key Program of Science and Technology of Anhui Province (No. 202203a07020008), Open Fund of Key Laboratory of Flight Techniques and Flight Safety, CACC (No, FZ2022KF15).

6. References

[1] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Information Fusion*, vol. 53, pp. 209–221, 2020.

[2] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," *A practical guide to sentiment analysis*, pp. 1–10, 2017.

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

[4] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[5] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Transactions on Affective Computing*, 2020.

[6] K. Somandepalli, T. Guha, V. R. Martinez, N. Kumar, H. Adam, and S. Narayanan, "Computational media intelligence: Human-centered machine analysis of media," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 891–910, 2021.

[7] L. Stappen, A. Baird, L. Schumann, and B. Schuller, "The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1334–1350, 2021.

[8] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *Ieee Access*, vol. 7, pp. 63 373–63 394, 2019.

[9] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2023.

[10] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 4730–4738.

[11] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2022.

[12] R. Lin and H. Hu, "Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, 2023.

[13] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3722–3729.

[14] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," *AI Open*, 2023.

[15] Y. Gu, X. Han, Z. Liu, and M. Huang, "Ppt: Pre-trained prompt tuning for few-shot learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8410–8423.

[16] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3045–3059.

[17] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[18] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[19] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

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

[23] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

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

[25] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 3718–3727.

[26] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[27] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2247–2256.

[28] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[29] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[30] R. Lin and H. Hu, "Multi-task momentum distillation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2023.

[31] F. Ma, Y. Zhang, and X. Sun, "Multimodal sentiment analysis with preferential fusion and distance-aware contrastive learning," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 1367–1372.

[32] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[33] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9180–9192.

[34] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.