Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Enhancing Multimodal Emotion Analysis through Fusion with EMT Model Based on BBFN

Wenshuo Wang*

School of Future Tech, South China University of Technology, Guangzhou, China

* Corresponding Author Email: 202364870251@scut.edu.cn

Abstract. Sentiment analysis, as one of the key technologies of natural language processing, has been widely used in medical, film and television fields. In order to increase sentiment analysis's precision, it is particularly important to integrate multi-modal data. This paper presents a pioneering fusion strategy that amalgamates the cutting-edge Efficient Multimodal Transformer (EMT) model with the innovative Bi-Bimodal Fusion Network (BBFN) to revolutionize emotion analysis. By synergistically integrating these two state-of-the-art models, the research endeavors to enhance the efficiency and precision of sentiment analysis in multimodal datasets by accentuating the intricate interplay of global-local cross-modal interactions. Through a rigorous process of meticulous experimentation and comprehensive analysis conducted on the challenging MOSI dataset, the integrated model unveils a plethora of groundbreaking advancements across pivotal metrics, including accuracy, correlation coefficient, and Mean Absolute Error (MAE). The innovative integration surpasses existing models and sets a new paradigm for multimodal sentiment analysis frameworks, highlighting the importance of holistic modal fusion in understanding human emotions.

Keywords: Multimodal sentiment analysis; multimodal fusion; cross-modal processing.

1. Introduction

With the goal of expanding sentiment analysis beyond text to include several modalities like photos, videos, and audio, multimodal sentiment analysis has become a crucial field in natural language processing and computer vision [1]. The current landscape witnesses significant efforts in integrating information from diverse modalities, leveraging advanced deep learning techniques, and addressing challenges in fusion strategies. Applications span diverse domains such as social media analysis, product reviews, and human-computer interaction, fueled by recent technological advancements and the growing availability of multimodal datasets. As research progresses, the field holds promise for deeper insights into human emotions across different modes of communication.

Nonetheless, a number of significant obstacles still exist in the field of multimodal sentiment analysis. Among these is the efficient combination of data from several modalities—text, image, video, and audio—to obtain a thorough comprehension of sentiment. Additionally, there's a need for robust feature extraction techniques capable of capturing nuanced emotional cues across different modalities. Another significant hurdle involves developing scalable and efficient algorithms that can handle the complex nature of multimodal data while maintaining high accuracy and performance. Apart from the challenges above, the lack of large-scale annotated multimodal datasets poses a considerable challenge for training and evaluating models in this domain. In order to advance multimodal sentiment analysis and realize its full potential in a range of applications, like market research, social media monitoring, and computer-human interaction, it is imperative that these challenges be resolved.

Scholars suggest that in order to tackle the difficulty of efficiently integrating heterogeneous data, it is necessary to extract and include significant information from various modalities while preserving their autonomy from one another [2]. Previous research in this area has mostly focused on early or late fusion techniques. Akhtar et al. developed a thorough framework for multi-task learning to simultaneously understand sentiment polarity and emotional intensity in a multimodal setting [3]. Rahman et al. combined functional gates and directly altered BERT to regulate the data flow between various modalities [4]. For sentiment analysis, Pham et al. suggested a cyclic translation method

between modalities to produce robust joint representations [5]. Besides models above, Wei H et al. proposed a novel fusion scheme called the Bi-Bimodal Fusion Network (BBFN) [2]. The model outperformed cutting-edge techniques on a number of measures, highlighting the significance of the introduced fusion scheme, regularization techniques, and control mechanisms in addressing the challenges of multimodal sentiment analysis.

Yet because the BBFN local-local cross-modal interaction modeling approach focuses exclusively on two pairings sharing a same core modality, it has a quadratic computing scale cost. This paper integrates Efficient Multimodal Transformer (EMT) model into the modal fusion portion of BBFN in order to solve this problem and improve the model's overall performance. The introduction of global multimodal interaction, which can successfully carry out global-local cross-modal interaction, is the fundamental component of EMT. Each interaction updates both global and local information, allowing it to have linear complexity and improve operational efficiency in contrast to previous local-local cross-modal interaction modeling systems that had quadratic computing scale costs [6]. In addition, this paper analyzes the performance of the new model on MOSI dataset [7].

2. Preliminary

2.1. Dataset

A well-known benchmark for assessing how well fusion networks perform in sentiment intensity prediction is the CMU-MOSI dataset [7]. The YouTube video blogs that comprise this collection feature speakers sharing their thoughts on a range of subjects. It contains 2,199 utterance-video segments from 93 videos featuring 89 different narrators. Each part has a manually written number score that indicates the relative strength of a pleasant or negative emotion, ranging from -3 to +3. The dataset offers a wide variety of video footage with sentiment labels for training and evaluation, making it an invaluable tool for researchers and developers working on sentiment analysis tasks. The CMU-MOSI dataset presents an invaluable chance to advance research in sentiment analysis and related topics because to its large annotated data collection. Table 1 displays the data set.

Table 1. Sample Dataset

Video ID	Speaker ID	Segment ID	Sentiment Intensity Score
1	1	1	2.7
1	1	2	-1.5
1	2	1	0.3
1	2	2	-2.1
2	3	1	1.8
2	3	2	-0.9
2	4	1	-2.5
2	4	2	2.0
3	5	1	1.5
3	5	2	-1.0
3	6	1	-1.8
3	6	2	2.2

2.2. Indicator Selection and Description

Mult_acc_7: This metric refers to the multi-class accuracy with seven discrete categories, indicating the model's performance in classifying sentiment across multiple classes.
F1 score: Combining precision and recall, the F1 score offers a fair assessment of the model's effectiveness in binary or multi-class classification tasks.

Accuracy: As a measure of how well the model predicts things overall, accuracy shows what percentage of cases are correctly identified out of all instances.
MAE (Mean Absolute Error): MAE calculates the average error magnitude between values that are predicted and those that are actual, providing insight into the model's prediction accuracy. Here is the MAE formula.

$$MAE(X,h) = \frac{1}{m}\sum_{i=1}^{m}|h(x_i) - y_i| \quad (1)$$

Correlation Coefficient: The correlation coefficient provides an indication of the prediction accuracy and consistency of the model by evaluating the magnitude and direction of the linear relationship between the predicted and actual values.

3. Method

The integration of the EMT model into the BBFN model enhances global-local cross-modal interactions and improves computational efficiency, leading to superior performance in multimodal sentiment analysis tasks. This improved model demonstrates superior performance on the MOSI dataset compared to existing models.

3.1. Bi-Bimodal Fusion Network

For multimodal sentiment analysis, BBFN is a unique fusion architecture [2]. It introduces a fusion technique made up of two bi-modal fusion modules to handle the difficulty of balancing the contributions of various modality pairs. BBFN model, unlike traditional ternary symmetric fusion methods, focuses on the text-visual (TV) and text-acoustic (TA) pairs of text-related modalities. These pairs are used as inputs for the bimodal learning modules, which iteratively encourage modalities to complement each other's information through interactive learning. The model is built on stacked Transformers, known for their efficiency in multimodal learning. In addition, the BBFN model incorporates modality-specific feature space separators and gated control mechanisms to provide fine-grained control and prevent feature space collapse throughout the fusion process. Experimental results demonstrate that the BBFN model outperforms current state-of-the-art approaches in multimodal sentiment analysis tasks.

3.2. Efficient Multimodal Transformer

An innovative framework designed for reliable Multimodal Sentiment Analysis (MSA) is the EMT model [6]. The inefficiency of simulating cross-modal interactions in unaligned multimodal data and the sensitivity to random modality feature absence are the two main problems of MSA that it attempts to overcome. By utilizing utterance-level representations from each modality as a global multimodal context, EMT enables efficient cross-modal interactions by engaging with local unimodal elements. Global and local attribute development is encouraged by EMT, which aims to improve sentiment analysis's efficacy and efficiency in multimodal data. Through its innovative design and approach, EMT sets a new standard for multimodal sentiment analysis frameworks.

3.3. Integrated model

To further enhance the benefits and improvements this paper integrated the EMT model into the modality fusion part of the BBFN model. The EMT model's focus on global-local cross-modal interactions allows for more efficient exploration of interactions between different modalities. By

incorporating EMT's approach, which has linear computational complexity, into the BBFN model, we can potentially reduce the computational cost associated with conventional local-local cross-modal interaction modeling techniques. Because the global multimodal context of the EMT model can interact with the local unimodal elements of the BBFN model in an effective manner, this integration may result in enhanced performance in multimodal fusion tasks. Furthermore, by utilizing the EMT model's hierarchical parameter sharing, the BBFN model can perform better overall and be more robust when it comes to multimodal sentiment analysis tasks by improving parameter efficiency and simplifying model training. The improved model demonstrates superior performance on the MOSI dataset compared to existing models.

4. Experiment

4.1. Visualization Analysis

In order to observe the performance of the model more intuitively, this paper conducted visualization. The visualization results are shown in Figure 1. The outcome shows that performance measures have steadily improved throughout epochs. The trend of the MAE values is downward, suggesting that the model's predictions are approaching the values seen in reality. The correlation coefficient also shows an increasing trend, suggesting a stronger linear relationship between predicted and actual values. The accuracy metrics and F1 score show that the model can accurately categorize emotions in a variety of modalities. An improved balance between recall and precision is shown by an increasing F1 score, which results in more accurate emotion recognition. The accuracy values show the percentage of correctly classified instances, which is consistently high, indicating the model's effectiveness in recognizing emotions. The multimodal emotion detection model appears to be learning from the input data and getting better with more time, based on the overall results.

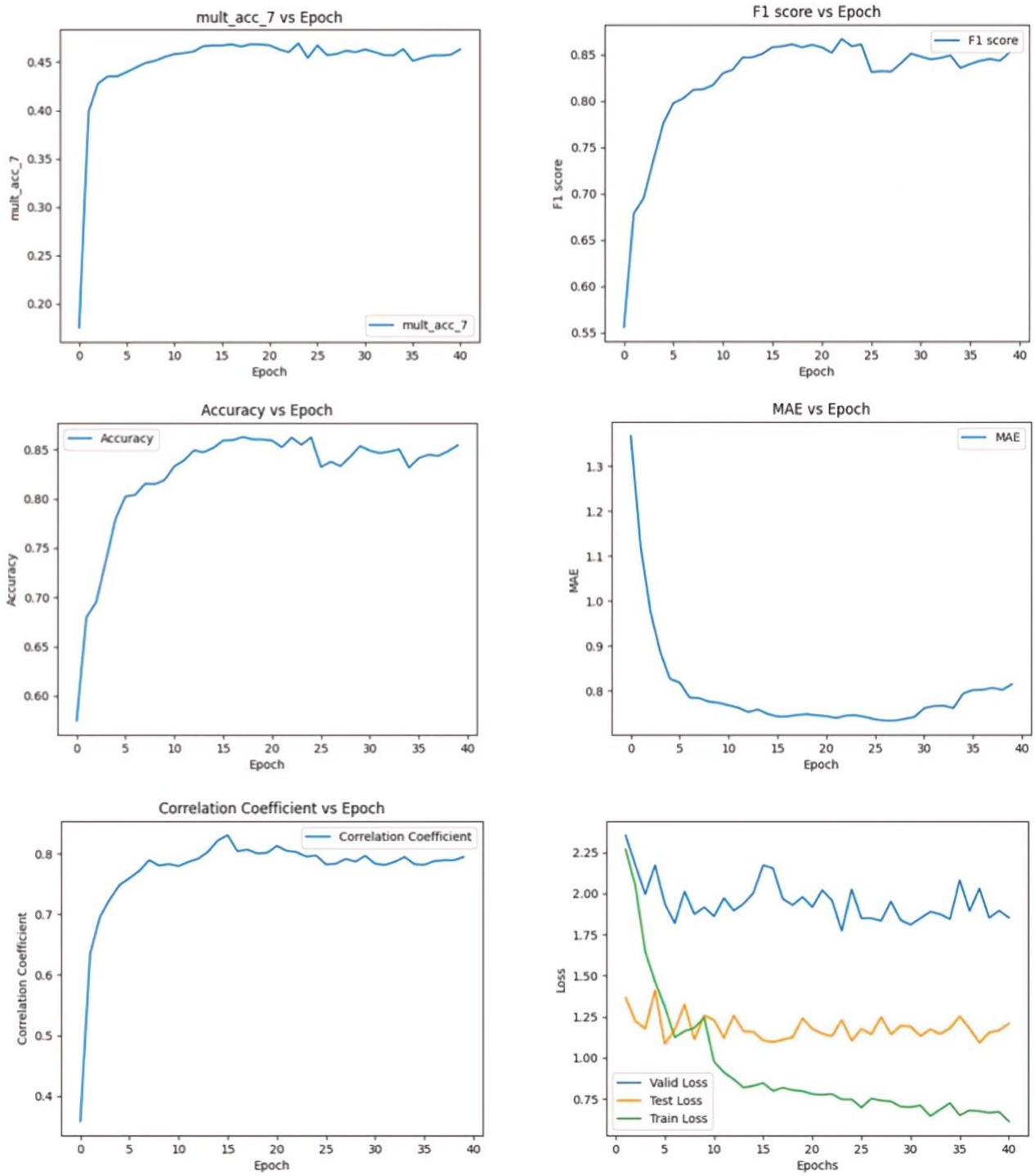

Figure. 1 Integrated model

4.2. Comparision to State-of-the-art

To assess the improvement of the model itself as well as the comparison with the SOTA model quantitatively, Table 2 is presented in this paper according to the evaluation metrics. This table compared the result of integrated model with top-performing MSA models: DFF-ATMF [8], Low-rank Matrix Fusion (LMF) [9], Tensor Fusion Network (TFN) [10], Multimodal Factorization Model (MFM) [11], Interaction Canonical Correlation Network (ICCN) [12], Multimodal Transformer (MulT) [13], Modality-Invariant and -Specific Representations (MISA) [14], Self-Supervised Multi-task Multimodal (Self-MM) [15], MultiModal InfoMax (MMIM) [16], Adaptive Multimodal Meta-Learning (AMML) [17], Transformer-based Feature Reconstruction Network (TFR-Net) [18], BBFN and EMT. According to the results, by integrating the EMT model into the modality fusion part of the BBFN model, the integrated model performs better across a number of metrics on the MOSI

dataset. Specifically, there are enhancements in MAE, Correlation Coefficient, Mult_acc_7, Mult_acc_2, and F1 score.

The improvements in MAE indicate that the integrated model has better accuracy in predicting sentiment labels. The increase in Correlation Coefficient suggests a stronger linear relationship between predicted and actual sentiment values. The enhancements in Mult_acc_7 and Mult_acc_2 metrics demonstrate improved accuracy in classifying sentiment across multiple classes. Additionally, the improvement in F1 score reflects a better balance between precision and recall in sentiment analysis tasks.

Overall, the integration of the EMT model into the BBFN model's modality fusion part has led to significant enhancements in performance across various evaluation metrics on the MOSI dataset, showcasing the effectiveness of leveraging global-local cross-modal interactions and hierarchical parameter sharing in improving MSA tasks.

Table 2. Results on the MOSI dataset

CMU-MOSI					
Models	MAE	Corr	Acc-7	Acc-2	F1
DFF-ATMF	-	-	-	80.9	81.2
LMF	0.917	0.695	33.2	82.5	82.4
TFN	0.901	0.698	34.9	80.8	80.7
MFM	0.877	0.706	35.4	81.7	81.6
ICCN	0.862	0.714	39.0	83.0	83.0
Mult	0.832	0.745	40.1	83.3	82.9
MISA	0.817	0.748	41.4	82.1	82
Self-MM	0.712	0.795	45.8	82.5	82.7
MMIM	0.700	0.800	46.7	84.1	84
AMML	0.723	0.792	46.3	-	-
TFR-Net	0.754	0.783	-	-	-
BBFN	0.776	0.775	45	84.3	84.3
EMT	0.705	0.798	47.4	83.3	83.2
NEWMODEL	**0.744**	**0.802**	**46.9**	**86.9**	**86.8**

The integration of the EMT model into the modality fusion part of the BBFN model has resulted in performance improvements on the MOSI dataset, possibly driven by several key principles:
Global-Local Cross-Modal Interactions: Leveraging the EMT model's focus on global-local cross-modal interactions enables a thorough exploration of relationships among diverse modalities. By embedding this principle within the BBFN model, the integrated approach better captures the intricate interactions between modalities, enhancing the accuracy of sentiment analysis outcomes.
Hierarchical Parameter Sharing: The hierarchical parameter sharing mechanism inherent in the EMT model facilitates efficient parameter utilization and sharing across various model levels. Integration of this mechanism into the BBFN model optimizes parameter efficiency, streamlines model training, and ultimately enhances performance in sentiment analysis tasks.
Linear Computational Complexity: The linear computational complexity of the EMT model ensures computational efficiency, decreasing the overall computational burden of the integrated model. This efficiency translates to faster inference and training times, bolstering the model's scalability and practicality for real-world applications.
By incorporating these principles within the modality fusion component of the BBFN model, the integrated model benefits from heightened global context modeling, improved parameter efficiency, and reduced computational complexity. These enhancements are reflected in superior performance on sentiment analysis tasks, as evidenced by improvements in metrics such as MAE, Correlation Coefficient, Mult_acc_7, Mult_acc_2, and F1 score on the MOSI dataset.

5. Conclusion

In the pursuit of advancing multimodal emotion analysis, the integration of the EMT model with the BBFN model has yielded promising results and opened new avenues for research in sentiment analysis. By harnessing the power of global-local cross-modal interactions, the integrated model demonstrates a profound understanding of emotions across diverse modalities, leading to substantial improvements in sentiment analysis tasks. The visualization analysis provides compelling evidence of the model's learning efficacy and performance enhancement over time, reinforcing the importance of comprehensive modal fusion in capturing the nuances of human emotions. This study not only contributes to the evolution of multimodal sentiment analysis frameworks but also underscores the transformative potential of integrating diverse modalities for a nuanced understanding of human emotions in real-world applications.

Going forward, the field of multimodal emotion analysis has a number of important topics that need further research and development. The creation of more complex fusion mechanisms that can efficiently capture and integrate data from a larger range of modalities, including text, visual, aural, and maybe other sensory inputs, is one possible direction for future research. Additionally, by offering insights into decision-making processes and boosting system confidence, improving the integrated models' interpretability and explainability could further advance the discipline.

Moreover, much work needs to be done so as to answer the concerns of managing multimodal data that is not aligned and lacking modality information. Strong techniques must be developed in order to solve these issues and improve the generality and general resilience of multimodal sentiment analysis models for use in real-world applications.

Furthermore, investigating the incorporation of sophisticated machine learning methodologies, like meta-learning, self-supervised learning, or reinforcement learning, may improve the versatility and efficacy of multimodal emotion analysis models. To guarantee the appropriate and moral application of these technologies, it is also essential to look into the ethical ramifications of applying multimodal emotion analysis in a variety of contexts, including social media, healthcare, and education.

In order to realize the full potential of multimodal emotion analysis models for comprehending and interpreting human emotions across a range of modalities, future research endeavors should concentrate on enhancing these models' sophistication, robustness, interpretability, and ethical issues.

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41: 423-443.

[2] Wei Han, Hui Chen, Alexander Gelbukh, et al. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. ICMI '21: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021.

[3] Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, et al. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1: 370–379.

[4] Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, et al. Integrating multimodal information in large pretrained transformers. In Proceedings of the conference Association for Computational Linguistics Meeting, 2020, NIH Public Access: 2359.

[5] Hai Pham, Paul Pu Liang, Thomas Manzini, et al. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the AAAI Conference on Artificial Intelligence,2019, 33: 6892–6899.

[6] Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao. Efficient Multimodal Transformer with Dual-Level Feature Restoration for Robust Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing, 2023, 1-17.

[7] Amir Zadeh, Rowan Zellers, Eli Pincus, Louis-Philippe Morency. Morency Multimodal sentiment intensity analysis in videos: Facial gestures and verbalmessages. IEEE Intelligent Systems, 2016, 31(6): 82–88.

[8] Feiyang Chen, Ziqian Luo, Yanyan Xu, Dengfeng Ke. Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis. arXiv:1904.08138, 2019.

[9] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, et al. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, 1: 2247–2256.

[10] Amir Zadeh, Minghai Chen, Soujanya Poria, et al. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, 1103–1114.

[11] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, et al. Learning Factorized Multimodal Representations. In International Conference on Representation Learning, 2019, 6558–6569.

[12] Zhongkai Sun, Prathusha Sarma, William Sethares, Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34: 8992–8999.

[13] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, et al. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[14] Devamanyu Hazarika, Roger Zimmermann, Soujanya Poria. MISA: Modality-Invariant and-Specific Representations for Multimodal Sentiment Analysis. In Proceedings of the 28th ACM International Conference on Multimedia, 2020, 1122–1131.

[15] Wenmeng Yu, Hua Xu, Ziqi Yuan, Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35: 10 790–10 797.

[16] Wei Han, Hui Chen, Soujanya 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, 9180–9192.

[17] Ya Sun, Sijie Mai, Haifeng Hu. Learning to learn better unimodal representations via adaptive multimodal meta-learning. IEEE Transactions on Affective Computing, 2022.

[18] Ziqi Yuan, Wei Li, Hua Xu, Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In Proceedings of the 29th ACM International Conference on Multimedia, 2021, 4400–4407.