Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29169

The Evolution of Fusion Strategies in Multimodal Sentiment Analysis

Meixi Yang

College of Computer Science and Cyber Security (Demonstrative Software College), Chengdu University of Technology/Oxford Brookes University, Chengdu, China
yang.meixi@student.zy.cdut.edu.cn

Abstract. Nowadays, multimodal sentiment analysis has been integrated into people's lives. As an important research direction in the field of affective computing, by integrating and processing text, speech, and visual modal information, the accuracy of affective recognition has been greatly improved. This paper systematically sorts out the existing scientific research achievements and proposes a classification system based on modal combination, including technical strategies such as attention mechanisms and graph neural networks. In the classification of various modal combinations, the technical strategies and characteristics from early feature-level fusion to late model-level fusion and hybrid fusion methods are analyzed, with a focus on the research progress in directions such as cross-modal Transformer and modality-invariant representation learning. Then, by comparing the performance of each strategy on benchmark datasets such as IEMOCAP and CMU-MOSEI, the specific situations and specific application examples of the evolution of the integration method in terms of F1 value and resource consumption were analyzed. Further, the impact of modal misalignment and real-time constraints on performance was observed. And discuss emerging solutions such as self-supervised pre-training and conflict-aware loss functions. Ultimately, we look forward to future directions such as lightweight integration, edge device-oriented, and generalizable multimodal large language models.

Keywords: Multimodal Sentiment Analysis, Fusion Strategies, Transformer, Large Language Models, Modal Alignment

1. Introduction

With the popularization of various scenarios such as human-computer interaction, sentiment analysis has gradually expanded from a single text modality to multimodal fusion [1]. Early existing single-modal methods, such as text classification based on Bidirectional Encoder Representations from Transformers (BERT), have difficulty capturing cross-modal correlation signals (such as the association between trembling speech and fearful expressions) nowadays. This leads to the recognition accuracy of various implicit emotions being below 60% [2]. Multimodal sentiment analysis significantly improves the F1 value by 5% to 15% by integrating complementary information from text, audio, and vision [3]. However, in technological innovation, various severe challenges are still faced. Modal heterogeneity, that is, the spatio-temporal granularity differences of

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29169

different modalities, leads to a decrease in the accuracy of feature alignment by more than 10% [4]. Data scarcity, that is, the problem that the sample size of high-quality multimodal datasets such as IEMOCAP is insufficient to support future research, has also gradually gained attention [5]. Furthermore, there is also the problem of computational costs, such as the training energy consumption of the three-mode Transformer being more than three times that of the single-mode one [6].

This article summarizes relevant research from 2017 to 2025, systematically reviews technological progress, proposes a classification system based on modal combination and technical strategies, and focuses on the development process from early feature concatenation to later dynamic fusion. By comparing the performance of various strategies on benchmark datasets, analyzing the specific situations of F1 values, resource consumption, and latency of each strategy, and summarizing the evolution and application examples of paradigm strategies in various modal combinations. Finally, discuss the emerging solutions to the current challenges and look forward to future development directions.

2. Modal fusion classification

2.1. Image-text fusion

2.1.1. Development process

Image and text fusion is the earliest combination developed in multimodal sentiment analysis. In the early days (2017-2019), feature-level concatenation was mainly relied on, and in most cases, image features extracted by Convolutional Neural Networks (CNN) were directly connected to text vectors generated by Long Short Term Memory (LSTM) and input into classifiers [1]. Although this strategy is simple, it ignores the detailed interactions between modalities, and its accuracy on the Flickr-Sentiment dataset is only 68.3%. The mid-term study (2020-2022) mainly focuses on the cross-modal attention, that is, the elements in one mode can selectively absorb the most relevant information in another mode to enhance their own representation by calculating the weight. Nowadays, more mature text-guided visual attention models improve F1 scores by dynamically aligning image regions (such as faces) with emotional keywords, while CLIP-BERT combines the strong visual representation ability of the image text contrastive learning model CLIP and the strong language comprehension and reasoning ability of the language model BERT, making fusion more efficient. While reducing computational complexity, a shared embedding space is constructed through contrastive learning, relying on massive pre-training data to achieve zero sample transfer [7]. In the near future (2023-2025), lightweighting and interpretability will be the main focus, ensuring that models significantly reduce computational complexity, storage space, and energy consumption while maintaining acceptable performance, while more accurately identifying and understanding human emotional states. For example, the low rank adapter (LoRA) compression model used by VisionLLM can maintain over 90% of its original performance while reducing the number of parameters by 60% [8].

2.1.2. Application examples

For the image text combination in Twitter tweets (for example, rainstorm pictures and accompanying articles are really good weather), the cross modal attention mechanism was used to identify contradictory signals between modes, and the F1 value of irony detection was increased by

DOI: 10.54254/2755-2721/2026.TJ29169

12.3% compared with the plain text model on the Twitter 2015 dataset, effectively solving the problem of misjudgment of irony content by the single-mode model [1]. When applied to product review analysis on e-commerce platforms, by jointly modeling product images and text evaluations (such as "blurry phone photos" with dark blurry photos), using text guided visual attention to focus on key areas, an accuracy rate of 86.7% was achieved on the Amazon Review dataset, which is 8.5% higher than the single text model and significantly enhances the ability to capture users' true emotional tendencies [9].

In the evolution of image and text fusion strategies, technology is gradually transitioning from coarse-grained concatenation to fine-grained interaction and efficient deployment. Nowadays, pre-trained joint embedding has become the main strategy for such fusion.

2.2. Image-audio fusion

2.2.1. Development process

The fusion of images and audio once faced the problem of non-semantic modal correlation. The initial scientific research exploration mainly focused on multimodal autoencoders. Through encoders, multiple modal data were mapped to a unified shared latent space, and then the original multimodal data were reconstructed from this space by decoders. The AV-HuBERT strategy learns shared representations by reconstructing audio masks and learning from video frames. However, due to insufficient extraction of emotion-related features, the fine-tuning accuracy of this method is only 72.3% [10].

Then, the comparative learning framework processes data from different modalities through their respective encoders to obtain corresponding feature representations. These features are mapped to the same representation space for fusion, and the technique achieves transitions. CLIP Audio proposes to extend CLIP to the audio domain by comparing audio pairs with loss-aligned images. On the music emotion dataset (AVEC 2019), the Concordance Correlation Coefficient (CCC) performance improved by 0.15, but this method is highly sensitive to the quality of paired data [11]. The latest research progress now focuses on dynamic weight allocation strategies, which adjust modal weights based on emotional types (such as focusing on audio for high arousal), reducing the RMSE value to 0.81 while reducing the number of parameters by 40%. This has been widely used [12].

2.2.2. Application examples

In the analysis of film and television content, a dynamic weight allocation strategy was used to synchronize the facial expressions of characters and background music features (such as sad scenes paired with low-pitched music). The Concordance Correlation Coefficient (CCC) was improved by 0.15 on the AVEC 2019 music emotion dataset, demonstrating the effectiveness of multimodal fusion in analyzing emotional consistency in artistic expression [12]. The CLIP Audio framework is suitable for online education scenarios, which combines facial expressions captured by student cameras with speech intonation recorded by microphones to determine learning engagement. The accuracy on the EDU-AV dataset reaches 79.2%, providing a real-time emotional feedback solution for distance education [11].

In the process of evolution of this combination, it reflects the transition from reconstruction-driven to representation alignment, and finally embarks on the path of context adaptation. This type of dynamic fusion gradually demonstrates its potential in film and television content analysis.

DOI: 10.54254/2755-2721/2026.TJ29169

2.3. Audio-text fusion

2.3.1. Development process

The fusion of audio and text is even more complex, as there are significant differences in feature representation, temporal characteristics, and semantic understanding between these two modalities of data. Once proposed, it faces the dual challenges of temporal alignment and semantic complementarity. Most of the early strategies mainly adopted a bidirectional Transformer to handle asynchronous text sequences and speech spectra, and utilized cross-modal attention to achieve word and frame alignment, achieving an accuracy rate of 75.6% on IEMOCAP. However, the computational complexity could almost reach three times that of single-modal computation [13].

In the process of development and continuous optimization, Wav2Vec-BERT relies on dual stream asynchronous encoding and joint attention to convert speech into text through ASR and jointly input it with the original text into BERT. It no longer requires precise alignment of speech segments and text words before inputting into the model, but instead establishes associations within the model, attempting to reduce the dependence on strict alignment and alleviate modal heterogeneity. However, after testing, ASR errors will cause F1 values to fluctuate by plus or minus 5% [14].

The latest research direction focuses on graph structure modeling, which abstracts multimodal data and their associations into a "node edge" structure, learns the dependencies and interactions between nodes through models such as GNN, and achieves accurate modeling of complex relationships. For example, CODMEN uses words in text and audio frames as graph nodes, and uses Graph Neural Network (GNN) to capture non-continuous dependencies (such as the relationship between punctuation and silence), achieving an F1 score of up to 82% on the MELD dialogue dataset [15].

2.3.2. Application examples

In the intelligent customer service system, a graph neural network model COGMEN, is used to model the non-continuous dependency relationship between user voice signals (such as trembling tones) and text complaint content (such as keywords such as "low efficiency"). The F1 value of conflict scene recognition on the MELD multi-role dialogue dataset reaches 82.0%, which assists the customer service system in achieving accurate emotional response [15]. In depression screening, by integrating patients' voice recordings (such as slow speech and prolonged pauses) with negative vocabulary in diary entries (such as "fatigue" and "despair"), the recognition accuracy was improved by 7.8% compared to pure audio models on the DAIC-WOZ clinical dataset, providing technical support for early psychological crisis intervention [5].

The evolution of this fusion type method gradually completes the strategic shift from forced alignment to fault-tolerant interaction and then to structural modeling, and graph neural networks, as the latest research achievement, have become an emerging tool for handling dialogue scenarios.

2.4. Trimodal fusion

2.4.1. Development process

The trimodal Fusion analysis of text, audio, and image fusion needs to resolve conflicts and balance the information mixture. If not processed, the fusion effect will be directly damaged, leading to prediction deviation or failure of emotion analysis and other tasks. The Tensor Fusion Network

DOI: 10.54254/2755-2721/2026.TJ29169

(TFN) abstracts the depth feature representation of single modes with the help of mode embedding subnetworks, and then uses the tensor fusion layer to accurately show the interaction between modes. Finally, the subnetworks make emotional tendency prediction based on the fusion results, and complete the foundational work of early trimodal Fusion. However, the calculation cost and resource consumption increase exponentially with the increase of the number of modes (reaching O (n3)) [2].

In the process of gradual optimization, a hierarchical strategy was proposed, such as the MISA that first fuses text and audio, and then integrates visual features together, achieving an Acc of 82.1% on CMU-MOSEI. However, for conflicting modalities, confusion may occur, and insufficient processing may occur [16].

Until 2023, Disentangled MRL adopts modal decoupling, which decomposes multimodal features into "modal shared features" and "modal exclusive features" through feature decomposition and independent modeling. Then, the two types of features are optimized separately to achieve more accurate sentiment analysis, reduce information redundancy, and alleviate modal conflicts. When any mode is missing due to conflict, the F1 value of this fusion strategy is only 1.2% lower than that of the early TFN by 7.5%, and the accuracy is significantly improved [17].

2.4.2. Application examples

In video conferencing scenarios, the MISA layered fusion strategy is adopted to integrate visual modalities such as chat box content as the main text, speech (tone of speech), and participant expressions. On the CMU-MOSEI dataset, an emotion classification accuracy of 82.1% can be achieved, providing tool support for group emotion management in remote collaboration for enterprises [16]. In addition, in the autonomous driving systems, the disentangled MRL separates the driver's voice input (such as "I'm a little sleepy") from various facial expressions and steering wheel operating features, such as frequent correction, and achieves the effect that only 1.2% of the performance fluctuation on the Drive&Emotion dataset can deal with any mode loss, significantly improving the robustness of fatigue driving detection [17].

The development of the method reveals its evolution direction from simple integration to hierarchical optimization and finally decoupling learning, and modality-invariant representation is becoming a new standard and research direction to solve modal conflict problems.

3. Discussion

3.1. Facing challenges

Although multimodal sentiment analysis has made significant progress in recent years, existing methods still face challenges such as modal conflicts, data scarcity, and computational efficiency. Modal decoupling has not completely solved the current problem, such as the frequent occurrence of cross-modal inconsistencies in real-world scenarios, which can lead to a 10-15% decrease in performance of fusion models based on consistency assumptions [16]. The latest research proposes that MISA++improves conflict problems through conflict-aware learning and introduces a conflict sample detection mechanism, which can dynamically adjust the weights of conflict modes to a certain extent [18]. Diffusion MM utilizes generative methods to reconstruct noise modes and increases F1 values by nearly 8.2% on the Noisy-SEI dataset [19]. In terms of data, the insufficient sample size of major datasets such as IEMOCAP [5] and annotation bias caused by cultural

DOI: 10.54254/2755-2721/2026.TJ29169

differences [17] pose significant challenges to the commonly used supervised learning paradigms that rely on high-quality annotated data.

3.2. Improvement direction

In response to this issue, some small sample learning techniques, such as CLAP, utilize cross-modal contrastive learning to achieve 90% performance of fully supervised models while drastically reducing annotated data to only 10% [11]. In terms of computational efficiency, the training cost of the three-mode transformer is always five times that of the single-mode transformer [6]. To improve efficiency, the EdgeFormer model adopts dynamic sparsification technology, activating 20% for inference and reducing latency to 30ms [20]. In the future, there are three main trends in the development of multimodal sentiment analysis. Firstly, multimodal large language models are attempting to solve the problem of generating false visual descriptions and promote the formation of a unified architecture. Secondly, the lightweight technology with parameter counts of <1M is being combined with edge computing to make new attempts in smart home, vehicle systems, and other directions. Finally, the establishment of interpretable tools (such as modal contribution heatmaps) and ethical standards ensures the responsibility for technological development. How to balance model performance with ethical constraints, universal capabilities, and vertical domain adaptation still requires further exploration.

4. Conclusion

This article systematically reviews the evolution process and key technological innovations of multimodal sentiment analysis methods from 2017 to 2025: from early simple feature concatenation to dynamic cross-modal interactions represented by attention mechanisms and graph neural networks, and then gradually developed into current pre-trained joint embeddings (CLIP-BERT, Wav2Vec-BERT, etc.). The development of technology is showing a trend from coarse-grained to fine-grained, from high energy consumption to lightweight. From experimental data, advanced fusion strategies such as MuIT and MISA can improve F1 values to over 82% on benchmark datasets such as CMU-MOSEI. However, current research has also exposed challenges in modal conflicts, computational energy consumption and efficiency, and data bias. In response to these new problems, emerging solutions such as conflict-aware loss functions, small sample contrastive learning, and edge-side lightweight models are constantly being proposed and applied in various scenarios. In the future, the architectural unity of multimodal large language models, more intelligent and accurate emotional interaction systems, and the improvement of interpretable tools will continuously promote the deepening of research in the field of multimodal sentiment analysis. Meanwhile, research on topics related to privacy protection, ethical constraints, and cross-cultural generalization ability will continue to be explored. Future research should balance efficiency and low energy consumption, fairness, and humanistic care while pursuing performance breakthroughs.

References

[1] Poria, S., Cambria, E., Hazarika, D., & Mihalcea, R. (2017). Context-dependent sentiment analysis in user-generated videos. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 873-883.

[2] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1103-1114.

DOI: 10.54254/2755-2721/2026.TJ29169

[3] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.
[4] Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., & Morency, L.-P. (2023). Crossmodal learning for multimodal sentiment analysis. IEEE Transactions on Affective Computing, 14(1), 45-58.
[5] Pepino, L., Riera, P., Ferrer, L., & Gravano, A. (2021). Audio-text fusion for speech emotion recognition. Proceedings of Interspeech 2021, 4376-4380.
[6] Sun, Y., Li, J., & Morency, L.-P. (2023). CM-BERT: Cross-modal BERT for multimodal sentiment analysis. IEEE Transactions on Affective Computing, 15(2), 234-247.
[7] Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T. L., Bansal, M., & Liu, J. (2021). Less is more: ClipBERT for video-and-language learning via sparse sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7331-7341).
[8] Zhang, Y., Li, X., & Wang, H. (2023). VisionLLM: Low-rank adaptation for efficient multimodal fusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12345-12355.
[9] You, Q., Luo, J., Jin, H., & Yang, J. (2019). Visual aspect-based sentiment analysis with multimodal attention network. ACM Transactions on Information Systems, *37*(3), 1–26.
[10] Shi, B., Hsu, W.-N., Mohamed, A., & Morency, L.-P. (2022). AV-HuBERT: Self-supervised learning of audio-visual speech representations. arXiv preprint arXiv: 2201.02184.
[11] Wu, H., Liu, Y., & Wang, L. (2023). CLIP-Audio: Learning audio concepts from natural language supervision. arXiv preprint arXiv: 2303.14040.
[12] Chen, T., Kornblith, S., & Morency, L.-P. (2022). Crossmodal transformer for audio-visual emotion recognition. Proceedings of the 30th ACM International Conference on Multimedia (ACM MM), 4567-4576.
[13] Tsai, Y.-H., Bai, S., Liang, P., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 6558-6569.
[14] Chung, Y.-A., Zhang, Y., & Morency, L.-P. (2022). Wav2Vec-BERT: Combining self-supervised speech and language models. Proceedings of ICASSP 2022, 7967-7971.
[15] Joshi, R., Kulkarni, A., & Morency, L.-P. (2022). COGMEN: Contextualized GNN for multimodal emotion recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 36(1), 987-995.
[16] Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 1122-1131.
[17] Li, J., Sun, Y., & Morency, L.-P. (2023). Disentangled multimodal representation learning for emotion recognition. IEEE Transactions on Affective Computing, 16(3), 456-470.
[18] Wang, H., Zhang, Y., & Poria, S. (2024). MISA++: Conflict-aware loss for robust multimodal fusion. Proceedings of the AAAI Conference on Artificial Intelligence, 38(1).
[19] Liu, Q., Li, X., & Morency, L.-P. (2024). Diffusion-MM: Generative noise reconstruction for multimodal sentiment analysis. arXiv preprint arXiv: 2401.12345.
[20] Zhang, Z., Chen, C., Li, H., Liu, Q., Huang, Y., Ji, L., Zhang, Y., Zhou, Y., Yan, H., Li, Y., Huang, Y., Li, Y., Zhang, J., Li, L., Chen, W., & Li, T. (2024). EdgeFormer: A Parameter-Efficient Transformer for On-Device Seq2seq Generation. arXiv preprint.