Research Progress of Multimodal Emotion Recognition Fusion Strategy

Binghe Zhang *

School of College of Electronic Science and Engineering, JiLin University, Jilin, 130012, China

* Corresponding Author Email: zhangbh1924@mails.jlu.edu.cn

Abstract. Multimodal emotion recognition relies on the collaborative perception of text, speech, visual and physiological signals. In recent years, it has achieved rapid development in the fields of emotional computing, human-machine interaction and medical health. With the emergence of large models, cross-modal attention mechanisms and dynamic fusion strategies, traditional decision-level fusion and feature-level fusion are unable to meet the requirements of refined, real-time and open vocabulary recognition. This article reviews five cutting-edge fusion methods, namely model-level fusion, cross-modal attention, multimodal fusion centered on large language models, dynamic fusion driven by reinforcement learning, and lightweight models for deployment. The article discusses the core concepts, advantages and disadvantages, and applicable scenarios of each method. Finally, it identifies current challenges and outlines future research directions.

Keywords: Multimodal Sentiment Recognition; Fusion Strategies; Cross-modal Attention; LLM-Centric Fusion; Lightweight Models.

1. Introduction

Multimodal Emotion Recognition is core research in the fields of human-machine interaction, emotional computing and social robots, and exist enormous application potential in the fields of intelligent assistants, emotional companion robots and customer service. Human beings have varities of emotions in communication, and they can communicate through a variety of sensory channels such as sound, language, vision, and physiological signals in daily life. Traditional emotion recognition methods mostly rely on single modality. However human emotion expression is very complex. A single modality not capture the truth emotion correctly in normal and which is usually prone to cheated. Therefore, the research of multimodal emotion recognition is particularly important.

The core of multimodal emotion recognition research is to effectively integrate multiple heterogeneous data sources such as vision, hearing and text to achieve better robustness. In recent years, deep learning has shown strong performance in extracting advanced features automatically and processing complex data [1]. Deep learning models, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) are good at processing sequence data. Transformer model has made a major breakthrough in cross-modal attention mechanism and solved the problem of long-distance dependence [2]. These models have made great contributions to multi-modal data alignment.

Although multimodal emotion recognition has made great progress, it still faces many challenges. For example, how to cope with data imbalance and missing problems efficiently, how to make the modal learn by itself when it faces complex scenes. Additionally, with the powerful performance of large language model (LLM) in the field of natural language, how to integrate LLM with multimodal data to further improve the accuracy of emotion recognition is a significant research direction at present and in the future.

This research aims to systematically review and deeply analyze deep learning technology and fusion strategy in the current field of multimodal emotion recognition, and explore the potential contribution and developing respect of large language models in this field. This article will analyze the core contributions, methodological innovations, experimental evidence and limitations of existing research, and prospect the future research directions, aim to provide valuable insights and inspiration

for further developing in this field. Therefore, the fusion of original data and features is called feature-level fusion. The advantage of this fusion

2. Classification and Comparison of Fusion Strategies

2.1. Traditional fusion methods

Traditional multimodal fusion methods mainly include feature-level fusion, decision-level fusion and hybrid fusion. Feature-level Fusion combines the original data from different modalities or lower-level features in the extracting stage, and then deriving results through model inference. Therefore, the fusion of original data and features is called feature-level fusion. The advantage of this fusion lies in its ability to preserve fine-grained information between modalities, enabling the model to learn more complex cross-modal relationships. However, it requires a lot of data support, and its poor interpretability, so it is hard to understand its decision-making. Decision-Level Fusion combines the decision results, such as voting and weighted average, after each modality completes the decision alone. However, it will lose the underlying association, and its performance is highly dependent on the performance of single modal. Hybrid Fusion combines the advantages of feature-level and decision-level fusion. However, it has complex structure, difficult parameter adjustment and high technical requirements.

2.2. Novel fusion method

2.2.1 Model-level fusion

Do not number your paper: All manuscripts must be in English, also the table and figure texts, Model-level fusion is the natural evolution of traditional fusion methods. It integrates different modal information into a unified deep learning model architecture, rather than simply combines at the feature or decision level, making full use of the advantages of deep neural networks. Through model-level fusion, the data of each modality can interact at different network layers to achieve cross-modal information flow. Model-level fusion often involves complex network design and training strategies. For example, a shared framework can quickly train all networks, and the input consists of various observations such as bird-eye view, state, and front view, rather than traditional camera views [3].

2.2.2 Cross-modal attention

With the development of attention mechanism, cross-modal attention has gradually become the key to multimodal fusion. Cross-modal attention allows model to focus on the relevant parts of other modalities while processing the information of one modality. Transformer-based design has become the main stream, and the implementation methods include bidirectional cross-attention, recursive cross-attention, hierarchical attention pooling and so on. For example, in emotion recognition, cross-modal attention mechanisms can help the model focus on relevant regions of facial expressions when processing specific emotions. Meanwhile cross-modal attention mechanisms simultaneously paying attention to emotional vocabulary in text or specific frequency ranges in speech prosody to enhance the ability to capture inter-modal information [4]. Specifically, the classification of attention is illustrated in the figure 1 [4].

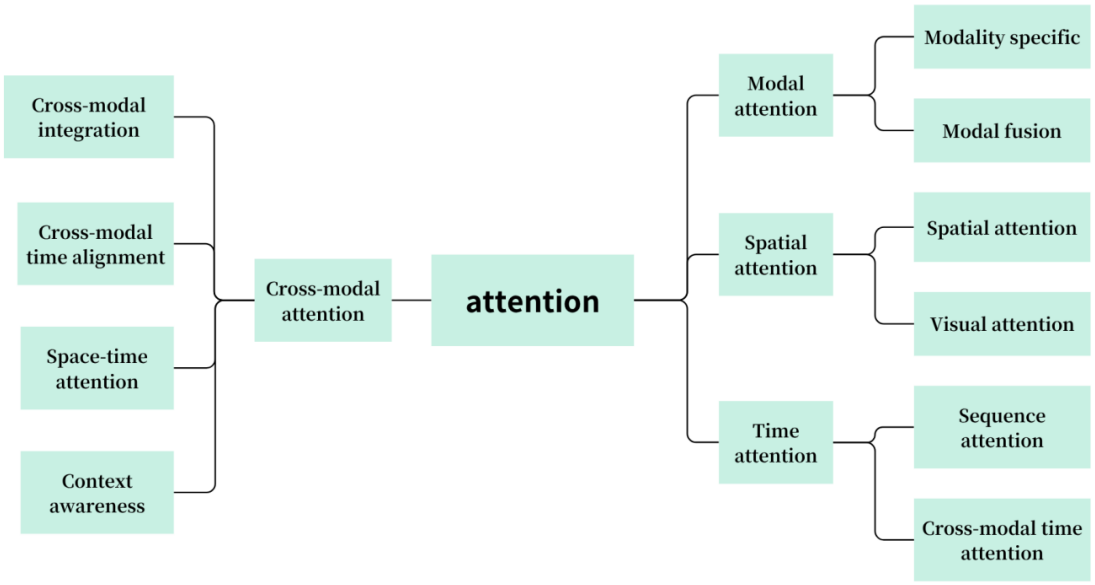

Fig. 1 Classification of Attention

2.2.3 LLM-Centric Fusion

In recent years, large language models (LLMs) have rapidly developed, and LLM-based fusion methods have gradually become a cutting-edge research direction. By feeding LLMs massive amounts of text data, they can perform self-supervised learning, thereby enabling them to master language skills and even perform complex logical reasoning. But language alone is not enough. Understanding a scene often requires combining multiple information such as images, sounds, and text to make judgments. Multimodal large language models (MLLMs) typically use a large language model (LLM) as the core of their system and build MLLMs by freezing or slightly fine-tuning a pretrained LLM. Freezing the LLM can maximize its knowledge retention and avoid forgetting during multimodal training. With the powerful text understanding and generation ability of LLM, MLLM can integrate and infer information from other modalities. For example, Magauiya Zhussip et al. proposed a framework across the Transformer layer, which reduces the attention module parameters by sharing dictionary atoms, thereby reducing the large demand for LLM computing and memory [5]. This also provides a more powerful semantic understanding ability for future cross-modal tasks.

2.2.4 Reinforcement learning-driven fusion

The basic algorithms of reinforcement learning include value iteration, policy iteration and dynamic programming. The core idea of these algorithms is to optimize action selection by calculating value function and policy. Next, the operation logic of value iteration, policy iteration and dynamic programming will be disassembled respectively.

The value iteration is a classical algorithm for solving Markov Decision Process (MDP). By iteratively updating the state value, the optimal strategy is found. First, set the value of all states to 0. For each state, the Q value (Q-Value) is calculated, and the Q value represents the expected cumulative reward after performing an action in a certain state [6, 7]. Formula (1) can be used to calculate

$$Q(s,a)=E[\sum_{t=0}^{\infty}\gamma^t r_t \,|s_0=s,a_0=a] \tag{1}$$

Among them, r_t is the reward of time t, γ is the discount factor $(0 \le \gamma \le 1)$. Finally, the Bellman equation of Formula (2) is used to update the state value.

$$V(s)=\sum_{a} P(s,a)\max_{a} Q(s,a) \tag{2}$$

Among them, P(s,a) is the probability of entering the next state after the action a is executed from the states. If the state value does not change within a certain range, the algorithm converges, otherwise the iteration continues [6, 7].

Policy iteration is an algorithm for solving Markov Decision Process (MDP), which iteratively updates the policy to find the best policy. First, the strategy of all actions is set to be evenly distributed, and then the Q value is calculated for each state using Formula (1). Finally, we use the formula (3) Softmax update strategy

$$\pi(a|s)=\frac{e^{Q(s,a)}}{\sum_{a'} e^{Q(s,a')}} \tag{3}$$

Among them, the Q value after the action a is executed from the states. If the strategy does not change within a certain range, the algorithm converges, otherwise it continues to iterate [6, 7].

Dynamic programming is a method to solve the optimal strategy of Markov Decision Process (MDP). Its core idea is to decompose the problem into a series of sub-problems and then solve them step by step [7].

Reinforcement learning (RL) is particularly suitable in complex environments, especially in multimodal data fusion. RL can adaptively adjust modal weights or fusion strategies through dynamic decision processes (such as MDP) to achieve closed-loop optimization of state-action-reward, so as to maximize long-term rewards. For example, in traffic signal control, deep reinforcement learning optimizes traffic signals by integrating the driving style of vehicles [8].

2.2.5 Lightweight deep learning fusion

In recent years, deep learning has developed rapidly, but the complexity of deep learning models is increasing, which makes the models unable to be deployed on devices with limited resources. Through model compression, knowledge distillation, pruning, more compact network structure, parameter sharing or quantization, lightweight deep learning fusion reduces the inference time and the consumption of computing resources, and reduces the demand for calculation and storage of the model, so that the model can be deployed on resource-constrained devices. This provides an advanced method for the application of emotion recognition on IoT devices [9]. For example, self-supervised emotion recognition based on skeleton data abstracts high-level semantics embedded in skeletal data [9].

3. Comparison of multimodal fusion

The five multimodal emotion recognition fusion methods described in this paper have their own advantages and disadvantages. Model-level fusion establishes cross-modal coupling in the network structure, cross-modal attention dynamically pays attention to the relevant information of other models by introducing attention mechanism, LLM-Centric fusion uses the semantic reasoning ability of large language models to enhance emotional understanding, reinforcement learning-driven fusion dynamically adjusts modal weights or strategies through reward mechanisms, the lightweight model maintains high performance under the condition of limited computing and storage resources. On the whole, there are differences between different methods in accuracy, efficiency and robustness. The specific comparison is shown in Table 1, and the performance comparison diagram is shown in Figure 2.

Table 1. Method comparison

Fusion Method Category	Core Idea	Advantages	Disadvantages	Application Scenarios
Model-level Fusion	Achieves subnetwork sharing or joint training at the model structure level.	Deep semantic interaction, captures cross-modal consistency.	Relatively high computational complexity, sensitive to missing modalities, high parameter requirements.	Video sentiment analysis, multimodal interaction systems.
Cross-modal Attention	Introduces attention mechanisms, enabling the model to focus on relevant parts of other modalities when processing one modality's information.	Effectively captures complex cross-modal relationships, enhances depth and efficiency of information integration.	Attention mechanism incurs high computational cost, large data volume and annotation requirements, higher quality demands.	Visual question answering, image captioning, cross-modal retrieval, vehicle recognition [10].
LLM-Centric Fusion	Maps non-text modalities into a unified semantic representation in language space, and performs reasoning with large language models (LLMs).	Utilizes the semantic reasoning power of LLMs, suitable for diverse and complex multimodal sentiment tasks.	Relies on large-scale pretrained models, slower inference speed.	Medical emotion monitoring, cross-lingual sentiment analysis.
RL-driven Fusion	Leverages reinforcement learning (RL) to dynamically optimize modality fusion strategies, learning optimal weights and fusion schemes through interaction.	Strong adaptability to dynamic and uncertain environments, can achieve adaptive and personalized fusion.	Training is complex and unstable, reward design is difficult, low sample efficiency.	Autonomous driving [3], human–robot interaction [11], adaptive control, gaming AI.
Lightweight Models	Focuses on efficiency and resource constraints through knowledge distillation, model compression, and efficient inference techniques.	Fast inference, low resource demand, suitable for edge and mobile devices.	Performance may be slightly lower than large-scale models, limited robustness in complex tasks.	Edge computing, mobile device applications, real-time interactive systems.

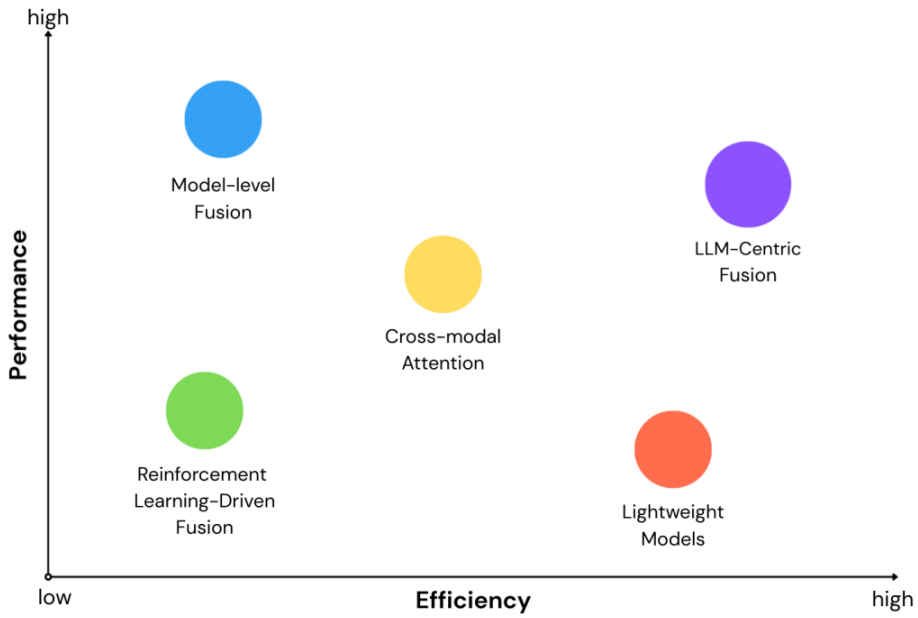

Fig. 2 Comparison of Fusion Strategies

4. Summary

Through research, this paper systematically reviews the evolution of multimodal emotion recognition fusion strategies and various technical frameworks, comparing and analyzing the advantages and disadvantages of different fusion strategies. We find that multimodal fusion strategies improve the robustness of recognition by fusing information. This paper reviews five cutting-edge strategies, namely model-level fusion, cross-modal attention, LLM-centric multimodal fusion, reinforcement learning-driven fusion, and lightweight models. Recent research indicates that multimodal emotion recognition is developing towards deeper cross-modality and increased deployment possibilities. To achieve more intelligent emotion recognition, further progress is needed in modal heterogeneity and alignment, dynamic fusion strategies, and lightweighting.

References

[1] Gajiwala, C. The Rise of Deep Learning and Neural Networks: Revolutionizing Artificial Intelligence. European Journal of Computer Science and Information Technology,2025.

[2] Chen, Z., & Zhang, B. Application of Multimodal Deep Learning Algorithm in Image and Text Fusion Recognition. 2025 IEEE 5th International Conference on Power, Electronics and Computer Applications, 2025, 599-604.

[3] C. Xu, H. Zhao, X. Lu, et al. A Deep Reinforcement Learning Method for Autonomous Driving Integrating Multi-Modal Fusion, IEEE Transactions on Intelligent Transportation Systems, 2025, 26(8): 11850-11863.

[4] Geetha A.V., Mala T., Priyanka D., et al. Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions. Information Fusion, 2024, 105.

[5] Magauiya Zhussip, Dmitriy Shopkhoev, Ammar Ali, et al. Share Your Attention: Transformer Weight Sharing via Matrix - based Dictionary Learning. arXiv, 2025, 2508.04581

[6] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, IEEE Transactions on Neural Networks, 1998, 9(5): 1054-1054

[7] Puterman, M. Markov Decision Processes: Discrete Stochastic Dynamic Programming. IMA Journal of Management Mathematics, 1994.

[8] T. Xu, Y. Pang, Y. Zhu, et al. Real-Time Driving Style Integration in Deep Reinforcement Learning for Traffic Signal Control, IEEE Transactions on Intelligent Transportation Systems,2025, 26(8): 11879-11892.

[9] Z. Zhang, F. Liang, W. Wang, et al. Skeleton-Based Pretraining with Discrete Labels for Emotion Recognition in IoT Environments, IEEE Internet of Things Journal, 2025, 12(15): 31856-31868.

[10] Ashutosh Holla B., Manohara Pai M.M., Ujjwal Verma, et al. MSFFT: Multi - Scale Feature Fusion Transformer for cross platform vehicle re - identification. Neurocomputing, 2024, 582.

[11] Lai, D., Zhang, Y., Liu, Y., et al. Deep Learning - Based Multi - Modal Fusion for Robust Robot Perception and Navigation. arXiv, 2025, 2504.19002.