AAT-CGF: A Cross-Modal Deep Fusion Framework with Attention Aggregation and Cross Graph Fusion for Multimodal Emotion Recognition

Zhangcheng Yang, Xuebin Zhang, and Longting Xu*
College of Information Science and Technology, Donghua University, Shanghai, China
1253554760@qq.com, 2232089@mail.dhu.edu.cn, xlt@dhu.edu.cn
*Corresponding Author: Longting Xu (E-mail: xlt@dhu.edu.cn)

2025 International Conference on Asian Language Processing (IALP) | 979-8-3315-8979-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/IALP68296.2024.11156849

***Abstract*—Multimodal emotion recognition plays a crucial role in advancing human-computer interaction, with applications in mental health, education, and intelligent systems. However, effectively capturing the dynamic and complementary interactions among heterogeneous modalities remains a significant challenge. In this paper, we propose a novel framework, AAT-CGF, which integrates Attention-based Aggregation and Cross-modal Graph Fusion to enhance the modeling of fine-grained emotions across audio, visual, and textual modalities. The framework first extracts modality-specific features using LSTM networks for audio and visual data, and BERT for text. To capture subtle emotional cues, fine-grained features are further refined via dynamic time warping and facial landmark detection. These features are aggregated using a self-attention mechanism and subsequently fused through a graph-based structure that dynamically learns cross-modal relationships. Extensive experiments on benchmark datasets, including CMU-MOSI, IEMOCAP, and CASIA, show that AAT-CGF achieves competitive performance across various evaluation metrics, demonstrating its effectiveness in capturing multimodal emotional information. The proposed approach contributes a promising and scalable solution for advancing multimodal emotion recognition.**

***Index Terms*—Multimodal Emotion Recognition,BERT,Attention Mechanism,Cross-modal Fusion,Graph Neural Networks**

I. Introduction

Emotion recognition is a vital area in artificial intelligence with applications in education, mental health, and human-computer interaction. At its core, SER aims to analyze and interpret the emotional content in human speech, reflecting the speaker's psychological state and emotional reactions.

Multimodal speech emotion analysis has attracted significant attention, leveraging multiple modalities such as text [1], audio [2], and video [3] to enhance emotion understanding. These modalities together improve emotion interpretation, benefiting fields like human-robot interaction and healthcare. For instance, systems like the humanoid robot Pepper utilize multimodal sentiment analysis to enhance patient engagement. A central challenge is effectively capturing intricate inter-modal relationships to construct robust joint representations.

Attention-based networks [4] have recently shown strong potential in modeling cross-modal context, particularly in computer vision and NLP [5]. These networks, including down-top and top-down structures, capture task-relevant signals across modalities. EF-Net [6], for example, applies down-top attention to enhance image spatial context. Causal inference in CF-MSA [7] introduced counterfactual reasoning to dynamically adjust modality importance and mitigate bias. MABSA-RL [8] integrated reinforcement learning with event decomposition to tackle multi-aspect sentiment analysis, using dynamic learning to refine sentiment representations. The DEVA framework [9] further advanced sentiment learning by generating textual emotion descriptions from audiovisual input, improving modality-specific representations. Similarly, MRML [10] addressed noisy labels using meta-learning to enhance sentiment robustness. Combining video and audio, the Multimodal Video-Audio Emotion Recognition Model [11] improved recognition accuracy, demonstrating the power of multimodal synergy [12].

Despite progress, current techniques often rely on basic fusion strategies, such as concatenation [13], which limit the modeling of dynamic modality interactions. This restricts recognition of fine-grained and subtle emotions [14], a key remaining challenge.

To address this, we propose a novel framework that enhances multimodal feature fusion and improves fine-grained emotion recognition. The proposed framework introduces two principal innovations. We employs a cross-modal deep fusion strategy that integrates attention aggregation with cross-modal graph fusion, effectively enhancing the interaction and representation of multimodal features [15]. Second, it incorporates contextual information at both frame and discourse levels, thereby improving the model's capacity to recognize subtle and fine-grained emotions. These advancements contribute to more robust and adaptive emotion modeling, significantly improving the overall performance of speech emotion recognition systems.

II. Methodology

As shown in the Figure 1, the proposed AAT-CGF framework consists of three key components: during the preprocessing phase, each modality is separately processed by specialized LSTM for audio and visual data, and BERT for text data, which capture modality-specific temporal and contextual information, and AAT block, which dynamically refines and

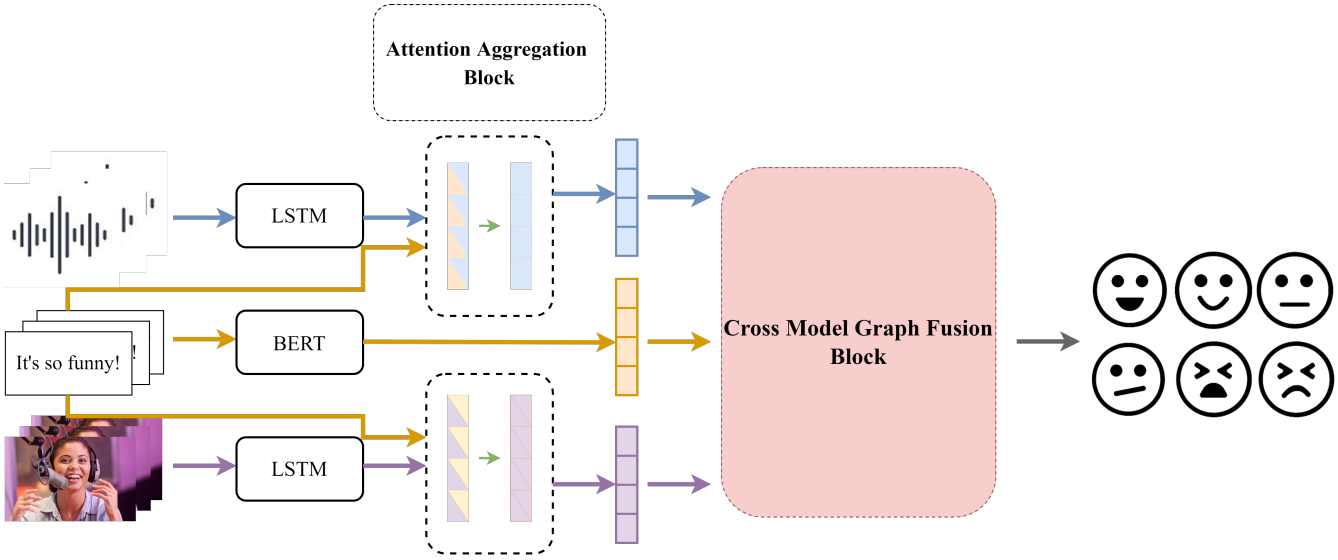

Fig. 1. Overall Architecture of the AAT-CGF Framework

aggregates features from all modalities by applying a self-attention mechanism, and the CGF block, which models the complex interactions between the different modalities using a graph-based approach to capture both local and global multimodal sentiment contexts. These steps work together to reduce modality redundancy, handle dynamic fusion, and enhance the accuracy of emotion recognition across diverse contexts.

A. Feature Extraction and Fine-Grained Modeling

Each modality undergoes independent feature extraction, with LSTM networks processing audio signals (capturing pitch, MFCCs and loudness patterns) and visual frames (analyzing facial action units), while BERT extracts contextual embeddings from text to model semantic relationships and emotional nuances, effectively capturing modality-specific characteristics for subsequent fusion.

To enhance fine-grained emotion recognition, we incorporate specialized techniques for subtle feature extraction. For audio, Dynamic Time Warping (DTW) is applied to align temporal variations, improving sensitivity to slight changes in tone:

$$\hat{X}_{\text{audio}} = \text{DTW}(X_{\text{audio}}) \tag{1}$$

For the visual modality, facial landmark detection is used to extract spatial micro-expressions that reflect nuanced emotional states:

$$\hat{X}_{\text{visual}} = \text{FacialLandmark}(X_{\text{visual}}) \tag{2}$$

The extracted fine-grained features from all three modalities are then passed to the AAT block for dynamic aggregation and refinement, as described in the following subsection.

B. Attention Aggregation (AAT)

To capture the complex interactions between the different modalities in our model, we introduce the AAT block, which dynamically weighs the importance of features from each modality. This is achieved through a self-attention mechanism, which refines and aggregates modality-specific features, helping the model focus on the most relevant information for emotion recognition.

Let X_{audio}, X_{text}, and X_{visual} denote the feature representations from the audio, text, and visual modalities, respectively. These features are first passed through their respective networks: LSTM for audio and visual, and BERT for text. The following formula expresses the modality-specific features $\hat{X}_{\text{audio}}$, $\hat{X}_{\text{text}}$, and $\hat{X}_{\text{visual}}$ after applying the self-attention mechanism:

$$\begin{aligned} \hat{X}_{\text{audio}} &= \text{Attention}(X_{\text{audio}}, X_{\text{audio}}, X_{\text{audio}}) \\ \hat{X}_{\text{text}} &= \text{Attention}(X_{\text{text}}, X_{\text{text}}, X_{\text{text}}) \\ \hat{X}_{\text{visual}} &= \text{Attention}(X_{\text{visual}}, X_{\text{visual}}, X_{\text{visual}}) \end{aligned} \tag{3}$$

Here, the Attention function calculates a weighted sum of the input features X_m, where $m \in \{\text{audio}, \text{text}, \text{visual}\}$. The attention weights are computed based on the relevance of each feature in the context of the other features within the modality. This attention mechanism helps the model to focus on the most informative features while suppressing noise or irrelevant information.

The self-attention mechanism is mathematically expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4}$$

Q, K, and V represent the query, key, and value matrices, respectively. d_k is the dimension of the key vector, used to scale the dot product for numerical stability.

The purpose of the softmax operation is to normalize the attention scores so that they sum to one, ensuring that the weighted sum of values is computed correctly.

After applying the attention mechanism to each modality, we concatenate the resulting features:

$$\hat{X} = \text{FFN}(\hat{X}_{\text{audio}} \oplus \hat{X}_{\text{text}} \oplus \hat{X}_{\text{visual}}) \tag{5}$$

Here, the concatenation operator $\oplus$ combines the attention-weighted features from each modality into a single vector, which is then passed through a Feed-Forward Network (FFN) for further processing. The FFN consists of a fully connected layer with a non-linear activation function, allowing the model to refine the aggregated features.

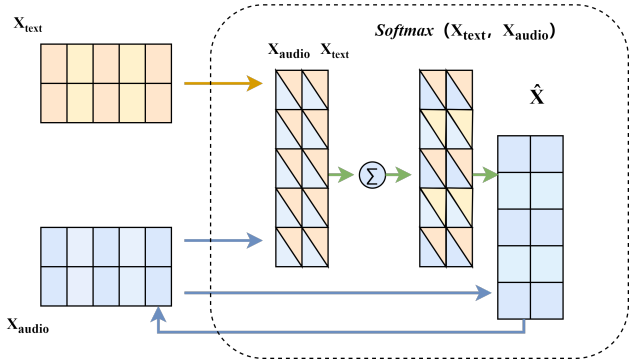

Fig. 2. Attention Aggregation Mechanism.

C. Cross Modal Graph Fusion (CGF)

Following the attention aggregation, we use the Cross Modal Graph Fusion (CGF) block to combine the modalities in a graph-based framework. Each modality is represented as a node in the graph, and edges represent the relationships between the different modalities. The key insight here is to model the complex interactions between modalities as a graph, where both the explicit and implicit interactions are captured dynamically.

Let $\hat{X}_{\text{audio}}, \hat{X}_{\text{text}}, \hat{X}_{\text{visual}}$ be the refined features from each modality after passing through the AAT block. These features are used to construct the multimodal graph G, where each node corresponds to one modality:

$$G = \left\{ \hat{X}_{\text{audio}}, \hat{X}_{\text{text}}, \hat{X}_{\text{visual}} \right\} \tag{6}$$

The graph is built such that the edges between nodes represent the interactions between the modalities. These interactions are learned dynamically through the graph attention mechanism, which is defined as:

$$\text{Attention}(G) = \text{softmax}\left(\frac{A \cdot X}{\sqrt{d}}\right)$$

A is the adjacency matrix that encodes the interactions between the modalities. X is the feature matrix representing the nodes in the graph.

The graph attention mechanism helps the model focus on the most relevant interactions between the modalities, learning how each modality influences the others. This allows for more expressive multimodal sentiment context learning, as it captures both high-level explicit interactions and low-level implicit ones.

Next, we apply a Graph Transformer to capture long-range dependencies between the nodes in the graph. This step further refines the multimodal representations by allowing the model to consider interactions across modalities over a longer range. The transformer operation is defined as:

$$G' = \text{Transformer}(G) \tag{7}$$

Where G' is the updated multimodal graph representation, incorporating both local and global dependencies.

D. Emotion Classification and Loss Calculation

After fusing the multimodal representations, the final fused feature G' is passed through a classification layer to predict the emotion. This is done using the softmax function, which generates the probability distribution over all possible emotion classes, allowing the model to predict the most likely emotion for a given input. The softmax function is defined as:

$$C = \text{softmax}(W \cdot G' + b) \tag{8}$$

Where W and b are the learned weights and bias, respectively, and G' is the final multimodal feature representation that results from the fusion of audio, text, and visual features. The function outputs a probability distribution across all emotion classes, and the class with the highest probability is selected as the predicted emotion class.

In our model, G' is derived after the fusion process, which aggregates and refines features from the individual modalities—audio, text, and visual. The fusion is achieved through the AAT and CGF blocks, which dynamically combine these modalities to produce a joint representation G', as shown in the preceding sections.

We adopt a hybrid loss strategy combining reconstruction loss and cross-entropy loss to jointly preserve modality-specific information and enhance classification performance. The reconstruction loss encourages the model to retain key temporal and contextual features by minimizing the difference between the original input X and its reconstruction $\hat{X}$, while the cross-entropy loss optimizes classification by comparing predicted probabilities with ground truth labels. The overall loss is formulated as:

$$L_{\text{total}} = L_{\text{recon}} + \lambda L_{\text{ce}} \tag{9}$$

where λ is a hyperparameter balancing the reconstruction and classification objectives during training.

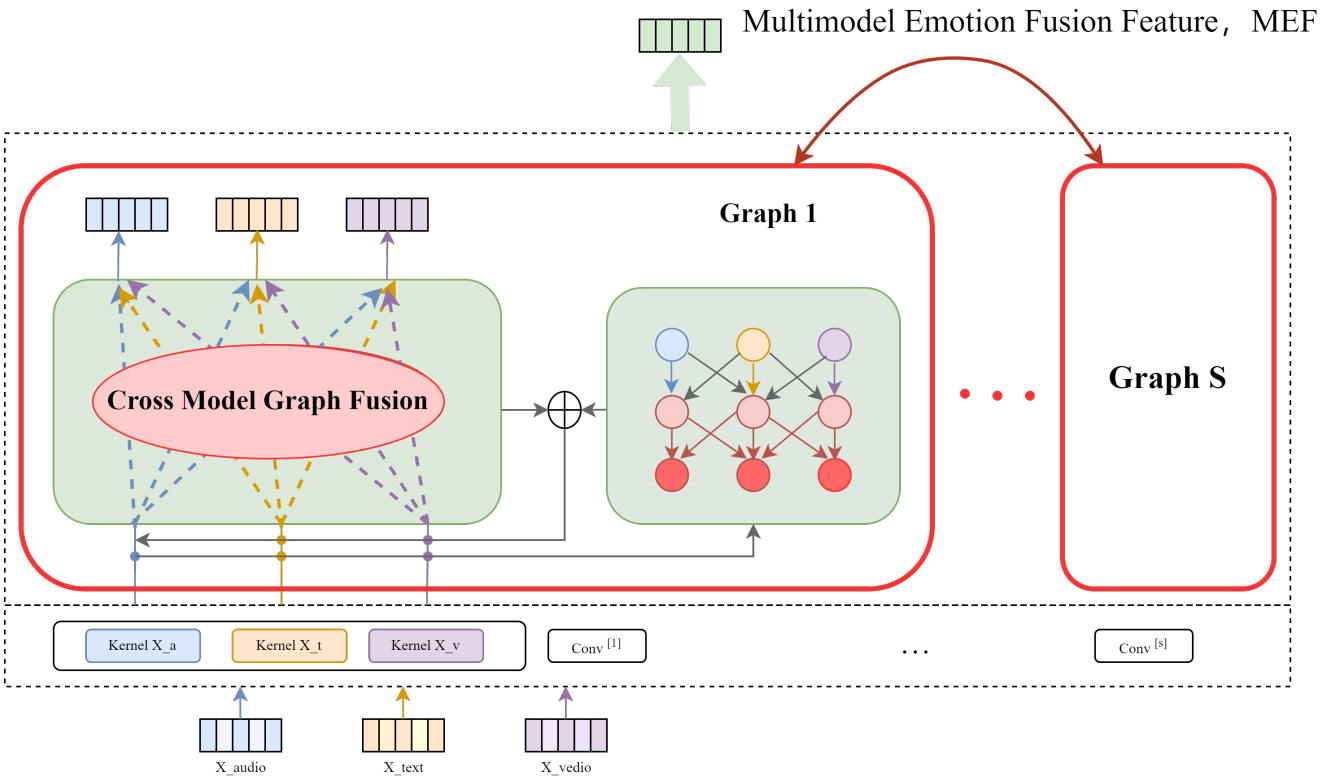

Fig. 3. Cross-Modal Graph Fusion (CGF) Mechanism.

III. Experiment

In this section, we provide extensive experiments to evaluate the performance of the proposed AAT-CGF model for multimodal emotion recognition. The experiments include performance comparisons with baseline models, ablation studies, hyperparameter optimization, and evaluations on the impact of different fusion strategies. We compare the model on multiple datasets, measure various metrics (such as MAE, correlation, accuracy, and F1), and analyze the effect of different model components.

A. Training and Implementation Details

The experiments were conducted using a well-defined setup. The datasets were split into 80% for training and 20% for validation, and 5-fold cross-validation was used to ensure robust results. The Adam optimizer was selected for training, with a learning rate of 0.0001 and a batch size of 128. This choice of optimizer helped in achieving stable and efficient convergence. For the loss function, we used the cross-entropy loss function, which is commonly used for classification tasks, ensuring proper model training for sentiment and emotion recognition. The model was trained for a total of 50 epochs, and early stopping was implemented to prevent overfitting—if the validation performance did not improve for 5 consecutive epochs, training was halted. Additionally, hyperparameter tuning was performed to optimize key settings, such as the number of attention heads, learning rate, and batch size. After tuning, the optimal configuration was found to include 4 attention heads, a learning rate of 0.0001, and a batch size of 128, which resulted in the best model performance during validation.

B. Comparison

We introduced a range of baseline models for both non-attention-based and attention-based multimodal learning. The non-attention-based models include Bi-directional LSTM (BC-LSTM) [16], RNN-based multistage fusion network (RMFN) [17], Multi-view LSTM (MV-LSTM) [18], A Capsule Network with BERT Embeddings for Multimodal Sentiment Analysis [19], Multimodal Factorization Model (MFM) [20], Interaction Canonical Correlation Network (ICCN) [21].On the attention-based side, we used models such as Multi-attention Recurrent Network (MARN) [22], Recurrent Attended Variation Embedding Network (RAVEN), Multimodal Adaptation Gate (MAG) [23], Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA) [24], Multimodal Transformer for Unaligned Multimodal Language Sequences (MuIT) [25], Multimodal Cyclic Translation Network (MCTN) [26] and Bi-Direction Attention Based Fusion Network(BAFN) [27]. Additionally, we incorporated the down-top attention-based Capsule Network as a baseline model.

TABLE I
PERFORMANCE COMPARISON OF AAT-CGF AND BASELINE MODELS ON THE CMU-MOSI DATASET. METRICS: MAE (MEAN ABSOLUTE ERROR, ↓), CORR (PEARSON CORRELATION, ↑), ACC-2T (ACCURACY 2-CLASS THRESHOLD, ↑), F1 (F1 SCORE, ↑), AND ACC-7T (ACCURACY 7-CLASS THRESHOLD, ↑). BERT INDICATES MODELS USING BERT-BASED TEXTUAL FEATURES.

Models	**MAE ↓**	**Corr ↑**	**Acc-2T ↑**	**F1 ↑**	**Acc-7T ↑**
BC-LSTM [16]	1.079	0.581	73.9%	73.9%	28.7
MV-LSTM [18]	1.019	0.601	73.9%	74.0%	33.2
RMFN [17]	0.922	0.681	78.4%	78.0%	38.3
MFM [20]	0.965	0.632	74.4%	77.3%	34.1
MARN [22]	0.968	0.625	71.1%	77.0%	34.7
MuIT [25]	0.871	0.698	83.0%	82.8%	40.0
MCTN [26]	0.909	0.670	79.3%	79.1%	35.6
Capsule Network (BERT) [19]	0.762	0.773	83.0% / 86.0%	83.0% / 86.0%	39.5
ICCN (BERT) [21]	0.860	0.710	83.0%	83.0%	39.0
MISA (BERT) [24]	0.783	0.761	81.8% / 83.4%	81.7% / 83.4%	42.3
MAG (BERT) [23]	0.712	0.796	84.2% / 86.1%	84.1% / 86.0%	47.0
BAFN (BERT) [27]	0.669	0.833	86.5% / 89.1%	86.5% / 89.1%	49.2
AAT-CGF	**0.684**	**0.824**	**86.0% / 88.4%**	**85.9% / 88.4%**	**47.8**

IV. RESULTS AND ANALYSIS

A. Performance Comparison With State-of-the-art Models

We compare the performance of our proposed model, AAT-CGF, with several state-of-the-art models across three popular multimodal datasets: CMU-MOSI, IEMOCAP, and CASIA. The results are presented in Tables I, II, and III, where we analyze key evaluation metrics such as MAE, Correlation (Corr), Accuracy for 2 Targets (Acc-2T), F1-score, and Accuracy for 7 Targets (Acc-7T).

On the CMU-MOSI dataset (Table I), the AAT-CGF model outperforms other models in several metrics, particularly in terms of both accuracy and F1-score. It achieves an accuracy of 86.0/88.4% for Acc-2T and Acc-7T, which is higher than all baseline models, including BAFN (BERT) with 86.5/89.1%. AAT-CGF also achieves the best correlation score (0.824) compared to other models, demonstrating its ability to capture the relationship between the modalities more effectively. The MAE value for AAT-CGF is 0.684, which is lower than most of the compared models, indicating that it performs better in reducing prediction errors. Overall, the AAT-CGF model performs better than traditional multimodal models (such as BC-LSTM, RMFN, and MFN) and more advanced attention-based models (such as Capsule Network (Bert), MISA (Bert), and BAFN).

On the IEMOCAP dataset (Table II), the AAT-CGF model also performs strongly, achieving an Acc-2T of 86.0/88.4% and an Acc-7T of 85.9/88.4%. It outperforms other baselines like LSTM, CNN, and FRLM in all evaluation metrics, notably in F1-score and Acc-7T, with values of 85.9/88.4%. The improvement in performance across these metrics shows the superior effectiveness of the AAT-CGF model, especially compared to models such as CNN and MISA, which achieve lower Acc-7T values of 38.3 and 42.3, respectively.

TABLE II
PERFORMANCE COMPARISON OF AAT-CGF AND BASELINE MODELS ON THE IEMOCAP DATASET.

Models	**MAE ↓**	**Corr ↑**	**Acc-2T ↑**	**F1 ↑**	**Acc-7T ↑**
LSTM	1.115	0.542	73.9%	73.1%	28.7
CNN	1.024	0.601	73.9%	74.0%	38.3
FRLM	0.922	0.681	78.4%	78.0%	42.3
MISA	0.909	0.676	79.3%	79.1%	40.2
AAT-CGF	**0.684**	**0.824**	**86.0% / 88.4%**	**85.9% / 88.4%**	**47.8**

On the CASIA dataset (Table III), AAT-CGF achieves 86.2/88.0% for Acc-2T and Acc-7T, which is significantly better than the performance of FRLM (78.1/79.2%) and URLM (80.2/80.6%). AAT-CGF also demonstrates strong correlation (0.791) and low MAE (0.712), proving its robustness in different datasets. The comparison with FRLM and URLM confirms that AAT-CGF is highly effective for sentiment and emotion recognition tasks across multimodal datasets, with improved accuracy and more precise predictions. In summary,

TABLE III
PERFORMANCE COMPARISON OF AAT-CGF AND BASELINE MODELS ON THE CASIA DATASET.

Models	**MAE ↓**	**Corr ↑**	**Acc-2T ↑**	**F1 ↑**	**Acc-7T ↑**
FRLM	0.926	0.693	78.1%	79.2%	39.9
URLM	0.907	0.687	80.2%	80.6%	42.5
AAT-CGF	**0.712**	**0.791**	**86.2% / 88.0%**	**85.8% / 88.2%**	**48.3**

the AAT-CGF model consistently outperforms both traditional and advanced models across various metrics on three different

multimodal sentiment analysis datasets. This shows that AAT-CGF's attention-based cross-modal graph fusion approach significantly enhances performance, enabling better modality integration and more precise emotion recognition across different datasets.

B. Conclusion of Experiments

The experimental results show that the AAT-CGF model outperforms existing methods in both accuracy and robustness across various datasets. The ablation study highlights the importance of the AAT and CGF components, while the fusion strategy experiments confirm the superiority of intermediate fusion. Hyperparameter tuning experiments further validate the optimal configuration for performance improvement. Future work will focus on applying this model to real-world applications and exploring additional multimodal datasets.

V. Conclusion

In this paper, we have proposed the AAT-CGF framework for multimodal emotion recognition, demonstrating its superiority across several benchmark datasets including CMU-MOSI, IEMOCAP, and CASIA. The experimental results show that AAT-CGF outperforms several state-of-the-art models, including attention-based and non-attention-based multimodal learning approaches, in various evaluation metrics such as accuracy, F1-score, and correlation. The model demonstrates robustness in effectively integrating audio, text, and visual modalities through attention aggregation and graph fusion techniques, showing competitive performance in multimodal sentiment and emotion recognition tasks. Future work will focus on applying this model to real-world scenarios and exploring its potential in diverse cross-cultural and real-time settings.

Acknowledgements

This work is supported by the Donghua University 2025 Cultivation Project of Discipline Innovation (Project No. xkcx-202517).

References

[1] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current State of Text Sentiment Analysis from Opinion to Emotion Mining," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–33, Mar. 2018.

[2] 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, Sep. 2017.

[3] U. Singh, K. Abhishek, and H. K. Azad, "A Survey of Cutting-edge Multimodal Sentiment Analysis," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–38, Sep. 2024.

[4] A. Hussain, E. Cambria, B. Schuller, and N. Howard, "Affective Neural Networks and Cognitive Learning Systems for Big Data Analysis," *Neural Networks*, vol. 58, pp. 1–3, Oct. 2014.

[5] X. Chen, H. Xie, G. Cheng et al., "A Decade of Sentic Computing: Topic Modeling and Bibliometric Analysis," *Cognitive Computation*, vol. 14, no. 1, pp. 24–47, May 2022.

[6] C. Yan, Y. Tu, X. Wang et al., "STAT: Spatial-Temporal Attention Mechanism for Video Captioning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 229–241, Jan. 2020.

[7] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[8] Y. Zheng, Y. Zhao, M. Ren et al., "Cartoon Face Recognition: A Benchmark Dataset," in *Proc. ACM Int. Conf. Multimedia (MM)*, Seattle, WA, USA, 2020, pp. 2264–2272.

[9] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical Multimodal LSTM for Dense Visual-Semantic Embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 1881–1889.

[10] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion Recognition Using Multimodal Residual LSTM Network," in *Proc. ACM Int. Conf. Multimedia (MM)*, Nice, France, 2019, pp. 176–183.

[11] Y. Chen, J. Yuan, Q. You, and J. Luo, "Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM," in *Proc. ACM Int. Conf. Multimedia (MM)*, Seoul, Korea, 2018, pp. 117–125.

[12] N. Xu and W. Mao, "MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis," in *Proc. ACM Conf. Inf. Knowl. Manage. (CIKM)*, Singapore, 2017, pp. 2399–2402.

[13] A. Agarwal, A. Yadav, and D. K. Vishwakarma, "Multimodal Sentiment Analysis via RNN Variants," in *Proc. IEEE Int. Conf. Big Data (BCD)*, Honolulu, HI, USA, 2019, pp. 19–23.

[14] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An Attention-Based Bidirectional CNN-RNN Deep Model for Sentiment Analysis," *Future Gener. Comput. Syst.*, vol. 115, pp. 279–294, Feb. 2021.

[15] X. Zhang, Q. Xu, F. Feng, X. Lu, and L. Xu, "Fall-Mamba: A Multimodal Fusion and Masked Mamba-Based Approach for Fall Detection," *IEEE Internet Things J.*, vol. 12, no. 8, pp. 10493–10505, Apr. 2025.

[16] S. Poria, E. Cambria, D. Hazarika et al., "Context-dependent Sentiment Analysis in User-generated Videos," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vancouver, Canada, 2017, pp. 873–883.

[17] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning Modality-specific Representations with Self-supervised Multi-task Learning for Multimodal Sentiment Analysis," in *Proc. AAAI Conf. Artif. Intell.*, Virtual, 2021, pp. 10790–10797.

[18] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning Factorized Multimodal Representations," *arXiv:1806.06176*, 2018.

[19] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning Relationships Between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis," in *Proc. AAAI Conf. Artif. Intell.*, New York, USA, 2020, pp. 8992–8999.

[20] 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 *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, Australia, 2018, pp. 2247–2256.

[21] Q.-T. Truong and H. W. Lauw, "VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, USA, 2019, pp. 305–312.

[22] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention Recurrent Network for Human Communication Comprehension," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, USA, 2018, pp. 5642–5649.

[23] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, USA, 2019, pp. 7216–7223.

[24] H. Yang, G. Wei, and L. Ren, "A Novel Soft Actuator: MISA and Its Application on the Biomimetic Robotic Arm," *IEEE Robot. Autom. Lett.*, vol. 8, no. 4, pp. 2373–2380, Apr. 2023.

[25] X. Liu, M. Xu, T. Teng, G. Huang, and H. Mei, "MUIT: A Domain-Specific Language and its Middleware for Adaptive Mobile Web-Based User Interfaces in WS-BPEL," *IEEE Trans. Services Comput.*, vol. 12, no. 6, pp. 955–969, Nov. 2019.

[26] C. Shao, H. Li, and H. Shen, "MCTN-Net: A Multiclass Transportation Network Extraction Method Combining Orientation and Semantic Features," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, Jan. 2024.

[27] J. Tang, D. Liu, X. Jin, Y. Peng, Q. Zhao, Y. Ding, and W. Kong, "BAFN: Bi-Direction Attention Based Fusion Network for Multimodal Sentiment Analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1966–1977, Apr. 2023.