ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240919001998
Original scinentific paper

Study on Multimodal Session Recognition Based on Graph Neural Networks

Zhixue WANG*, Hongwu ZHANG, Kai KANG

Abstract: As one of the special research directions of text emotion recognition task, the conversation emotion recognition task needs to take the conversation context, speaker and other factors into account to accurately identify the emotion state. To address challenges in multimodal information fusion and global-local feature extraction, this paper proposes a multimodal session recognition method based on multilevel attention mechanism and multistream graph neural network. Firstly, two key issues of the multimodal session recognition problem are analysed; secondly, a multimodal session emotion recognition model based on the improved attention mechanism strategy and the improved graph neural network is proposed to address the two issues; and finally, the effectiveness of the proposed method is verified through the analysis of simulation experiments. The simulation results show that the method achieves good accuracy in IEMOCAP and MOSEI datasets and effectively improves the efficiency of session fusion recognition.

Keywords: attention mechanisms; graph neural networks; information fusion; multimodal session recognition

1 INTRODUCTION

With the development of Internet technology, the transformation of the user's network use mode has improved the user's participation [1]. As more and more users share their personal views and experiences on the Internet, the conversation platform as one of the most important social ways, users express their own views through text, audio, video, audio and video converted to the form of this paper, mining text features, and recognition of text content [2]. Conversational emotion recognition task, as one of the special research directions of text emotion recognition task, needs to take into account the conversation context, speaker and other factors to accurately identify the emotional state [3]. The general conversational sentiment recognition task lacks a certain degree of reasoning ability, and needs to be analysed accurately with the help of external knowledge conversational sentiment [4]. Therefore, it is necessary to study the improvement of conversational sentiment recognition methods. Currently, the research on conversational sentiment recognition generally adopts textual sentiment recognition methods, which are divided into sentiment recognition methods based on sentiment dictionaries, sentiment recognition methods based on traditional machine learning, and sentiment recognition methods based on deep learning, depending on the method used [5]. Emotion recognition methods were based on emotion dictionaries when the emotion dictionary was constructed manually to match the emotion words in the text, so as to obtain the text emotion polarity [6]. Zhong et al [7] propose an extended dictionary-based sentiment recognition method, which performs sentiment analysis based on the existing sentiment dictionary, and extends the sentiment dictionary using linguistic rules and vocabulary to increase the coverage and accuracy of sentiment words. The sentiment recognition method based on the sentiment dictionary fails to take into account the contextual relationships of words as well as sentences, making the classification and recognition ineffective [8]. Sentiment recognition methods based on traditional machine learning can complete the sentiment recognition task with different classifiers [9]. Commonly used machine learning algorithms for emotion recognition methods based on traditional machine learning include K-nearest neighbours [10], plain Bayes [11], and support vector machines [12] methods. Hao et al [13] propose a multi-feature fusion method based on support vector machines to extract text features from four aspects such as word form, sentiment, syntax, and semantics. Sentiment recognition methods based on traditional machine learning are not effective in recognising non-linear and high latitude datasets because of the use of traditional machine learning algorithms [14]. Deep learning-based sentiment recognition methods make predictions based on context and do not rely on manually labelled corpus [15], and commonly used methods include convolutional neural networks, recurrent neural networks, long and short-term memory networks, and attention mechanisms [16]. Zhang et al [17] combine convolutional neural networks and long and short-term memory networks to construct a sentiment analysis model considering contextual relationships. Sentiment recognition methods based on deep learning mine the deeper meanings of data by learning the intrinsic laws of the data without manually designing features [18]. Session sentiment recognition targets the session itself to mine user sentiment information, which is mainly divided into context-based sentiment recognition and user information-based sentiment recognition [19]. The current research on the task of conversational emotion recognition only models from the traditional context, speaker and external knowledge, which makes the constructed model lack a certain degree of common sense perception and contextual reasoning ability [20].

Through the above literature analysis, the current multimodal conversational emotion recognition has the following two problems: 1) the fusion of different modal information suffers from the noise interference problem [21]; and 2) the global-local conversational emotion features are poorly captured [22]. In order to solve the multimodal information fusion problem and the global-local session emotion feature extraction problem, this paper proposes a multimodal session recognition method based on multilevel attention mechanism and multistream graph neural network. In this paper, by describing and analysing the multimodal information fusion problem and the global-local session sentiment feature extraction problem, we construct a multimodal session recognition

model based on improved graph neural network by combining the multilevel attention mechanism and multiflow graph neural network. The effectiveness and applicability of this paper's method are verified by public datasets, and the results show that the proposed method is feasible and efficient.

2 PROBLEM ANALYSIS

In order to improve the accuracy and robustness of the conversation emotion recognition model, the multimodal conversation recognition model is constructed by analysing the multimodal information fusion and global-local conversation emotion feature extraction, combining a variety of sensory modalities, and the specific model structure analysis is shown in Fig. 1.

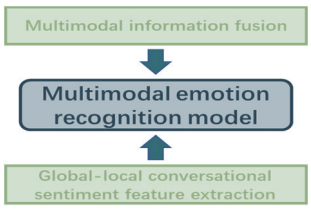


Figure 1 Structure of the multimodal problem analysis

2.1 Multimodal Information Fusion Problems

Considering multiple sensory modalities (facial expressions, speech, physiological signals and other modal information), this paper considers the problem of temporal alignment between different modalities [23]. In this paper, the multimodal representations of conversations are simply connected together by using the attention mechanism to extract contextual information and fused, as shown in Fig. 2.

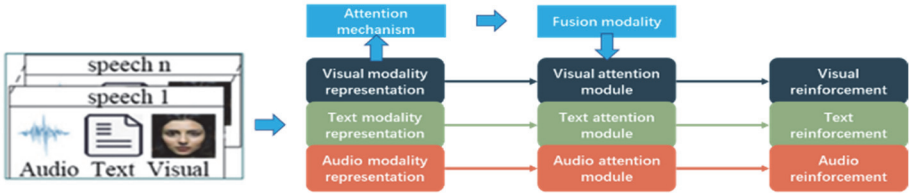


Figure 2 Diagram of the multimodal information fusion

2.2 Global-local Session Sentiment Feature Extraction Problem

In order to better model conversational emotion recognition, this paper extracts conversational emotion features from both global and local aspects [24].

Figure 3 Global-local session sentiment feature extraction fusion

Global information helps the model to understand the topic and context of the conversation, and can accurately capture the emotional tendency and attitude of the interlocutor [25]; local information helps the model to analyse the details and changes during the conversation, and can more accurately identify the emotional state and expression of the interlocutor [26]. In this paper, we first use the local emotional dependency between the interlocutors during the conversation, then capture the contextual information, fuse the outputs of each modal flow, and further fuse the global features of the modal ground, as shown in Fig. 3.

3 A MULTIMODAL SESSION RECOGNITION MODEL

3.1 Functional Analysis

Based on the analysis of the multimodal conversation recognition problem, this paper proposes a Multimodal Conversation Emotion Recognition Combining Multi-Level Attention and Multi -Stream Graph Neural Networks (MCER-MAMGNN) based on the Improved Attention Mechanism Strategy and Improved Graph Neural Networks. -Stream Graph Neural Networks, (MCER-MAMGNN). MCER-MAMGNN includes Multi-Level Attention Mechanism Strategy, Multi-Stream Graph Neural Networks Algorithm, and Multi-Stream Merge Classification Strategy, and the specific overall framework is schematically shown in Fig. 4.

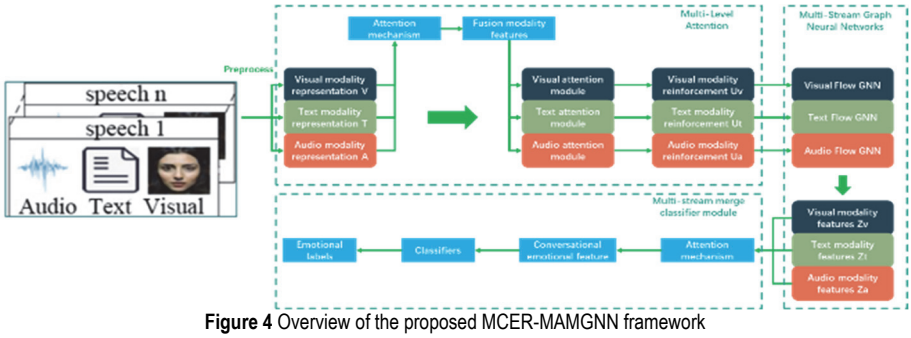


Figure 4 Overview of the proposed MCER-MAMGNN framework

As can be seen from Fig. 4, the multi-level attention mechanism strategy is mainly used for modal fusion of features and semantic interaction between video modality, text modality, and audio modality; the multi-stream graph

neural network algorithm is mainly used to extract the global-local session features of each modality; and the multi-stream merger classification strategy is mainly used for modal fusion and construction of classifiers.

3.2 Multi-level Attention Mechanism Strategy

In order to effectively fuse multimodal information, this paper adopts a multilevel attention mechanism strategy to enhance modal representation [27]. The specific structure of the multilevel attention mechanism strategy is shown in Fig. 4, which is mainly composed of the self-attention module and the cross-modal attention mechanism.

(1) Self-attention module.

Self-attention mechanism is mainly used for the first time to fuse the information of each modality, which is mainly composed of multi-head attention mechanism and feed-forward fully connected layer, and the specific structure is shown in Fig. 5. As can be seen from Fig. 5, audio modality, text modality and visual modality are spliced into fused modalities, and the session emotion modality features are obtained through the multi-head attention layer, normalisation layer, feed-forward fully connected layer and normalisation layer.

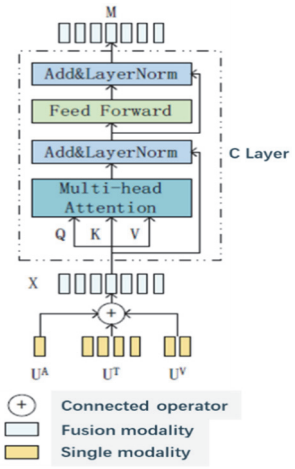


Figure 5 Multi-level attention module

Assuming that the session contains three modes, i.e., audio modal u_i^a , textual modal u_i^t , and visual modal u_i^v the three modes are spliced together to obtain the fusion modal u_i^{atv} :

$$u_i^{atv} = \left[u_i^a \oplus u_i^t \oplus u_i^v \right] \tag{1}$$

where $u_i^{atv} \in \mathbb{R}^{d_{atv}}$, $d_{atv} = d_a + d_t + d_v$. The fusion representation is dimensionally divided into r_1 copies, i.e., r_1 heads, and the ith head attention matrix is computed using scaled dot-product:

$$head_i = \text{softmax}\left(\frac{Q_i \left(K_i\right)^T}{\sqrt{d_k}} \right) V_i \tag{2}$$

where, $Q_i = XW_i^Q$, $K_i = XW_i^K$, $V_i = XW_i^V$ are three vectors generated from the fused modal representation after linear variation, i.e., query vector, robust vector, and value vector; X denotes the fused modal representation, i.e., $X = \left[u_1^{atv}, u_2^{atv}, ..., u_n^{atv} \right] \in \mathbb{R}^{n \times d_{atv}}$; $W_i^Q \in \mathbb{R}^{d_{atv} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{atv} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{atv} \times d_k}$ are three randomly initialised weight matrices; and $d_k = d_{atv} / r_1$.

The attention matrix obtained through the head is spliced to obtain the multi-head attention layer output, calculated as follows:

$$U' = \left[head_1 \oplus head_2 \oplus \cdots \oplus head_{r_1} \right] W^o \tag{3}$$

where W^o denotes the weight matrix, i.e., $W^o \in \mathbb{R}^{d_{atv} \times d_{atv}}$.

The normalisation layer, the feed-forward fully connected layer and the next normalisation layer are calculated as follows:

$$U = LN\left(X + U'; \gamma_1, \beta_1 \right) \tag{4}$$

$$M' = \text{Relu}\left(UW_1 \right) W_2 \tag{5}$$

$$M = LN\left(X + M'; \gamma_2, \beta_2 \right) \tag{6}$$

where, LN is the layer normalisation, Relu is the activation function, $\gamma_1 \in \mathbb{R}^{d_{atv}}$, $\beta_1 \in \mathbb{R}^{d_{atv}}$, $\gamma_2 \in \mathbb{R}^{d_{atv}}$, $\beta_2 \in \mathbb{R}^{d_{atv}}$ are the parameters, $W_1 \in \mathbb{R}^{d_{atv}}$, $W_2 \in \mathbb{R}^{d_{atv}}$ denote the trainable weight matrix of the feed-forward fully connected layer, $M = \left[m_1, m_2, ..., m_n \right]^T$ denotes the session fusion modal features containing contextual information, $M \in R^{n \times d_{atv}}$.

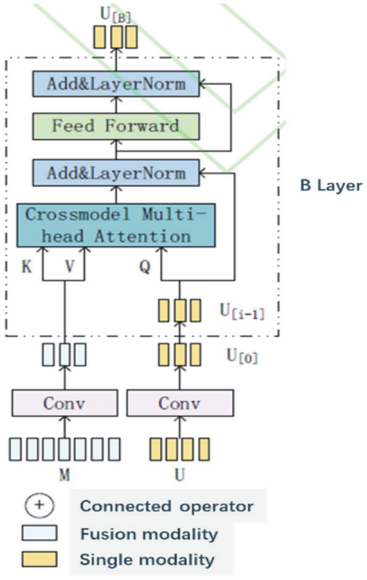


Figure 6 Cross-modal attention mechanism

(2) Cross-modal attention mechanisms.

The cross-modal attention mechanism mainly uses fused modal features to enhance the individual modal representations [28], which mainly consists of a cross-modal multi-head attention layer, and a feed-forward fully-connected layer, and the specific structure is shown in Fig. 6. The cross-modal attention mechanism adopts the Crossmodal Transformers model structure, with the robust vector and the value vector coming from the fused modal representation, and the query vector coming from each modal representation.

Combining each modality $u' \in \mathbb{R}^{n \times d}$ and the fused modal features $m' \in \mathbb{R}^{n \times d}$, the ith head attention matrix is calculated as follows:

$$head_i' = \text{softmax}\left(\frac{Q_i^U \left(K_i^M\right)^T}{\sqrt{d_{k'}}}\right) V_i^M \tag{7}$$

Among them, $Q_i^U = XW_i^{Q'}$ is the vector generated by the linear variation of each modal representation, i.e., query vector; $K_i^M = XW_i^{K'}$, $V_i^M = XW_i^{V'}$ are the vectors generated by the linear variation of the fusion modal representation, i.e., robust vector, value vector; $W_i^{Q'} \in \mathbb{R}^{d \times d_k}$, $W_i^{K'} \in \mathbb{R}^{d \times d_k}$, $W_i^{V'} \in \mathbb{R}^{d \times d_k}$ are the three randomly initialised weight matrices, $d_{k'} = d_{atv} / r_2$, r_2 are the number of heads.

The cross-modal multi-head attention layer for layer i is calculated as follows:

$$\begin{aligned} &CM_i^{M \to U}\left(U_{i-1}^{M \to U}, M\right) = \\ &= \left[head_1' \oplus head_2' \oplus \cdots \oplus head_{r_1}'\right] W'^o \end{aligned} \tag{8}$$

where $U_{i-1}^{M \to U}$ is the cross-modal attention block output for layer $i-1$ and $W'^o \in \mathbb{R}^{d \times d}$ is the weight matrix.

3.3 Multi-Stream Map Neural Network Algorithm

Considering the influence of different modal global-local features on the session recognition model, this paper adopts a multi-stream graph neural network [29] to carry out global-local session feature modelling for each modal representation separately, as shown in Fig. 7. From Fig. 7, the multi-stream graph neural network consists of multiple modal streams with the same structure, and each modal stream only processes specific modal data. The multi-stream graph neural network algorithm includes a graph construction module, a graph transformation network module, a graph convolutional network module, a graph transformer module, and a bidirectional MOGLSTM module.

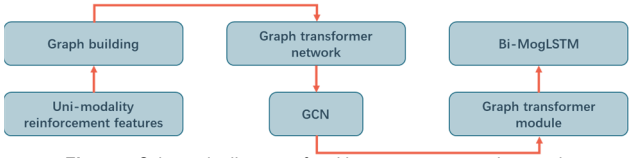


Figure 7 Schematic diagram of multi-stream map neural network

(1) Graph building blocks.

In order to obtain the discourse relations between the session participants and the session participants themselves, this method uses a graph building module to construct the session relations spatially and temporally. The relationship between the same session participants is represented as follows:

$$\begin{aligned} IR\left(u_i^{S_1}\right) = \Big\{ &u_i^{S_1} \leftarrow u_{i-P}^{S_1}, \cdots, u_i^{S_1} \leftarrow u_{i-1}^{S_1}, \\ &u_i^{S_1} \leftarrow u_i^{S_1}, u_i^{S_1} \to u_{i+1}^{S_1}, \cdots, u_i^{S_1} \to u_{i+F}^{S_1} \Big\} \end{aligned} \tag{9}$$

where $U^{S_i} = \left(u_1^{S_i}, u_2^{S_i}, ..., u_n^{S_i}\right)$ denotes the set of speakers i of the session and $U = \left\{U^{S_1}, U^{S_1}, ..., U^{S_T}\right\}$ denotes the session.

The relationship between the speeches of the different interlocutors is expressed as follows:

$$\begin{aligned} ER\left(u_i^{S_1}\right) = \Big\{ &u_i^{S_1} \leftarrow u_{i-P}^{S_2}, ..., u_i^{S_1} \leftarrow u_{i-1}^{S_2}, \\ &u_i^{S_1} \to u_{i+1}^{S_2}, ..., u_i^{S_1} \to u_{i+F}^{S_2} \Big\} \end{aligned} \tag{10}$$

where $\leftarrow$ denotes past relationships and $\to$ denotes future relationships.

(2) Graph Transformation Module.

In order to model the differences of session emotional dependence in different modalities and capture the node relationship heterogeneity, this paper adopts graph transformation neural network to construct the relationship heterogeneity between nodes. The graph transform neural network consists of D graph transform layers, and generates the neighbourhood matrix by learning the graph structure, which is calculated as follows:

$$Q_i^t = \varphi\left(A^{t-1}, softmax\left(W_i^{\varphi}\right)\right) \tag{11}$$

$$A_i^t = Q_{i1}^t Q_{i2}^t \tag{12}$$

where Q_i^t denotes the graph structure, A^{t-1} denotes the set of adjacency matrices, A_i^t is the new set of adjacency matrices, φ is the convolution operation, and $W_i^{\varphi} \in \mathbb{R}^{1 \times 1 \times N}$ is the parameter of φ.

(3) Graph Convolutional Networks Modules.

In order to capture the relationships between sessions, graph convolutional network is used to obtain different importance of different session relationships and from that a new set of weighted neighbourhood matrices is obtained. The graph convolutional network [30] module makes use of each adjacency matrix and accumulates the convolution results under different relationship graphs, which is calculated as follows:

$$P = \sum_i^N w_i \,\text{Re}\,\text{lu}\left(GCN\left(U, A_i^D\right)\right) \tag{13}$$

where w_i denotes the weighting operation, Relu denotes the activation function, GCN denotes the convolution operation and N is the number of matrices.

(4) Figure converter module.

In order to obtain local features, this paper adopts Graph Transformer model with homography learning and attention coefficients to obtain feature representations. The specific information fusion representation is as follows:

$$g_i = W_1 p_i + \sum_{j \in \tau(i)} a_{i,j} W_2 p_j \tag{14}$$

$$a_{i,j}^c = softmax\left(\frac{\left(W_3 p_i\right)^T \left(W_4 p_j\right)}{\sqrt{d}} \right) \tag{15}$$

where, $P = \{p_1, p_2, ..., p_n\}$ denotes the node features obtained by GCN, W_1 ,W_2, W_3 and W_4 are the coefficient matrices, $a_{i,j}$ denotes the attention coefficients, computed from the multi-head dot product attention, $a_{i,j}^c$ is the cth head attention matrix, and d is the head hiding size.

(5) Bidirectional MOGLSTM module.

In order to deal with the dependency of new input and hidden state directly pressed, this paper adopts BiMogLSTM method to construct context information and enhance the capability of modelling context information. BiMogLSTM neural network is an improved version of LSTM, which uses the gate mechanism to interact the current moment input g with the hidden state h_{prev} to get the new input and the hidden state, which makes the new input relevant to the context. BiMogLSTM neural network is calculated as follows:

$$\overrightarrow{G} = \overrightarrow{MOGLSTM}\left(G\right) \tag{16}$$

$$\overleftarrow{G} = \overleftarrow{MOGLSTM}\left(G\right) \tag{17}$$

where $\leftarrow$ denotes inverse and $\rightarrow$ denotes forward. The output of BiMogLSTM is obtained by splicing the bidirectional output:

$$Z = \{z_1, z_2, ..., z_n\} \tag{18}$$

3.4 Multi-stream Merger Classification Strategy

In order to improve the overall recognition performance and efficiency, this paper proposes a multi-stream merge classification strategy to merge multiple streams in everything. The strategy first splices the outputs of each modal stream together, and then adopts the self-attention mechanism to learn the importance degree of each modality, and finally outputs the corresponding session emotion categories. The specific computation of the multi-stream merge classification strategy is as follows:

$$y_i = \text{argmax}\left(\text{softmax}\left(W_2 \text{ReLU}\left(W_1 h_i^{'} + b_1\right) + b_2\right)\right) \tag{19}$$

where W_1, W_2 denote the coefficient matrix, b_1, b_2 denote the bias, y_i denotes the prediction type of the session u_i, and $h_i^{'} = \left[z_i^a \oplus z_i^t \oplus z_i^v \right]$ denotes the fusion sentiment feature.

3.5 Multimodal Session Recognition Method Flow Steps

Combining the multi-level attention mechanism, multi-stream graph neural network, multi-stream merger classification strategy, this paper proposes a multi-modal session recognition method based on improved graph neural network; the method flow chart is shown in Fig. 8, and the specific steps are as follows:

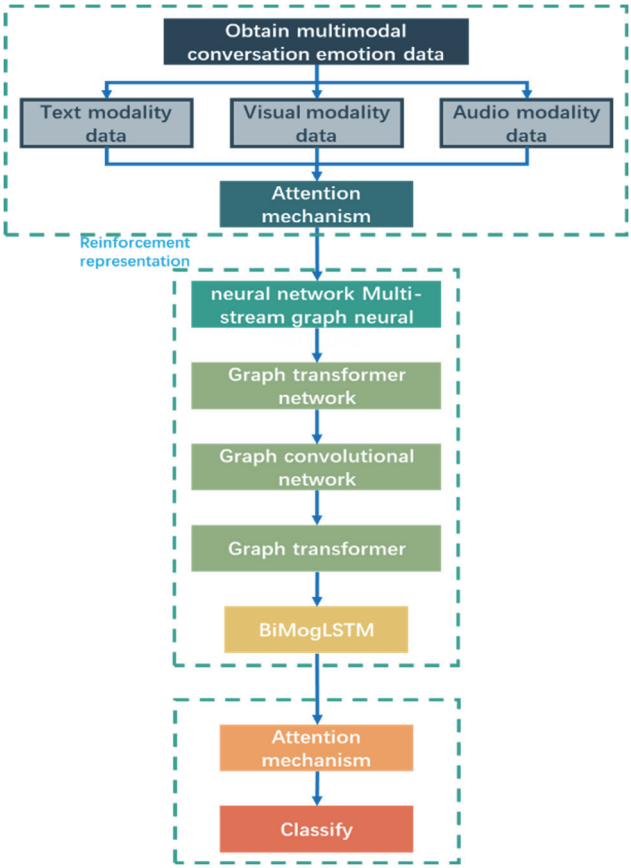


Figure 8 Flowchart of the multimodal session recognition method based on improved graph neural network

- Step 1: Acquire session data through sensors;
- Step2: Input the multimodal session representations such as speech, text, and vision into the attention mechanism module to obtain the fused modal features;
- Step 3: The first multimodal information fusion is achieved by the cross-modal attention module, which combines the modal representations and obtains the enhanced representation of each modality;
- Step 4: Input the reinforcement representation of each modality into the multistream graph neural network to construct the conversation relationship graph with each discourse as a node;
- Step 5: Construct session relationship heterogeneity features using GTN network and local features using GCN and Graph Transformor;
- Step 6: Capture contextual information using BiMogLSTM neural network;

- Step 7: Fuse the outputs of each modal stream using the self-attention module to extract contextualised session sentiment features and complete the second multimodal information fusion;
- Step 8: Use the classifier to complete the session recognition task.

4 EXPERIMENTS AND ANALYSIS OF RESULTS

In order to validate the effectiveness and efficiency of the proposed multimodal session recognition method based on improved graph neural network, the learning and training results of the proposed algorithm are analysed and discussed in this section by selecting the IEMOCAP dataset and the MOSEI dataset.

4.1 Data Description

(1) IEMOCAP data set.

IEMOCAP is a binary multimodal emotion recognition dataset containing six emotion categories such as anger, excitement, happiness, frustration, and neutrality in each session. OpenSmile is used for feature extraction of session speech, and the features can be reduced to 100 dimensions by normalisation and fully connected layer; Openface is used for face image and retrograde feature extraction, and the features are obtained to be 512 dimensions; SBERT is used to extract the text features of the session, and the semantic representations are extracted, and the features are obtained to be 768 dimensions.

(2) MOSEI data set.

MOSEI utilises the largest dataset for sentiment analysis and emotion recognition of online video sessions, where each session contains the categories of happy, sad, angry, fear, disgust, and surprise. Attention mechanism and weighting are used to obtain 160-dimensional audio features and 70-dimensional visual features, similarly, 768-dimensional session text features are extracted using SBERT.

(3) Data division.

The IEMOCAP dataset and the MOSEI dataset are divided into training, validation, and testing sets, as shown in Tab. 1.

Table 1 Division of data

data set	training set	validation set	test set
IEMOCAP	5810		162
MOSEI	16327	1871	4662

4.2 Experimental Configuration

(1) Experimental environment configuration

The experimental simulation environment is Windows 10 with 2.80GHz CPU, 8GB RAM and programming language Python 3.8.

(2) Comparison of algorithm settings

The Adamw optimiser is used in this experiment with batch size set dimension 32, number of iterations dimension 65 and initialised learning rate dimension 0.0003.

4.3 Evaluation Indicators

In this experiment, Accuracy and F1-score were used as evaluation indexes.

(1) Accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

where TP, TN, FP, FN denote the number of correct samples with correct model predictions, the number of incorrect samples with correct model predictions, the number of correct samples with incorrect model predictions, and the number of incorrect samples with incorrect model predictions, respectively.

(2) $F1$ value ($F1$-score).

$$F1 - score = \frac{2 \times (P \times R)}{P + R} \tag{21}$$

$$R = \frac{TP}{TP + FN} \tag{22}$$

$$P = \frac{TP}{TP + FP} \tag{23}$$

where R denotes the model recall and P denotes the model precision.

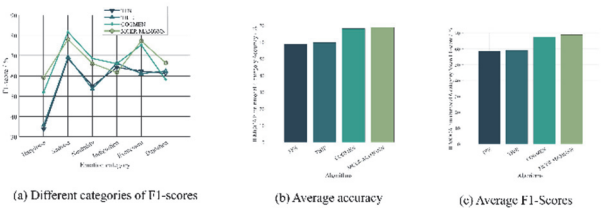
4.4 Analysis of Experimental Results

In order to verify the effectiveness and superiority of the multimodal session recognition method based on improved graph neural network, MCER-MAMGNN is compared with other three models such as TFN, TBJE, COGMEN, etc., and the evaluation results of each model are shown in Fig. 9, Fig. 10 and Fig. 11.

The results of the IEMOCAP multimodal 6-category emotion comparison experiment are given in Fig. 9. From Fig. 9, under the 6-category setting, MCER-MAMGNN has $F1$-scores values of 59.1%, 78.0%, 66.1%, 61.6%, 77.2%, and 66.4% under the happy, sad, neutral, angry, excited, and frustrated emotion categories, respectively, and its average accuracy is 69% and average $F1$ is 68.9%; both the average accuracy and the average $F1$, MCER-MAMGNN ranked first, and the other algorithms ranked COGMEN, TBJE, and TFN in that order.

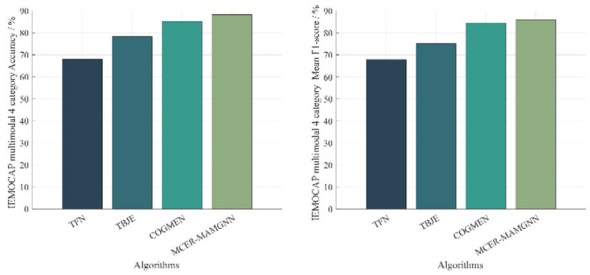
Fig. 10 gives the results of the IEMOCAP multimodal 4-class emotion comparison experiment. As can be seen from Fig. 10, under the 4-class setting, MCER-MAMGNN has an average accuracy of 88.3% and an average $F1$ of 86.0%; MCER-MAMGNN ranks first in both the average accuracy and the average F1, and the other algorithms rank, in order, COGMEN, TBJE, and TFN.

Fig. 10 gives the results of the IEMOCAP multimodal 4-class emotion comparison experiment. As can be seen from Fig. 10, under the 4-class setting, MCER-MAMGNN has an average accuracy of 88.3% and an average $F1$ of 86.0%; MCER-MAMGNN ranks first in both the average accuracy and the average F1, and the other algorithms rank, in order, COGMEN, TBJE, and TFN.



(a) Different categories of F1-scores (b) Average accuracy (c) Average F1-Scores

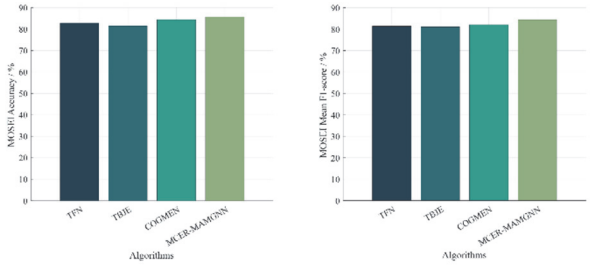
Figure 9 Results of the IEMOCAP multimodal 6-category emotion comparison experiment



(a) Average accuracy (b) Average F1-Scores

Figure 10 Results of the IEMOCAP multimodal 4-category emotion comparison experiment

Fig. 11 demonstrates the results of the MOSEI multimodal sentiment classification comparison experiment. As can be seen from Fig. 11, the average accuracy of MCER-MAMGNN, COGMEN, TBJE, and TFN is 85.70%, 84.43%, 81.50%, and 82.88%, respectively, and the average F1 is 84.45%, 82.12%, 81.1%, and 81.4%, respectively; in terms of the average accuracy, the rankings are, in order, MCER-MAMGNN , COGMEN, TBJE, and TFN; in terms of average F1, the rankings were, in order, MCER-MAMGNN, COGMEN, TFN, and TBJE.



(a) Average accuracy (b) Average F1-Scores

Figure 11 MOSEI multimodal sentiment classification comparison experiment results

5 CONCLUSION

In this paper, a multimodal session emotion recognition model based on improved attention mechanism strategy and improved graph neural network is proposed for two key points of the multimodal session recognition problem. The method analyses the multimodal information fusion problem and the global-local session emotion feature extraction problem; for the multimodal information fusion problem, multilevel attention mechanism and cross-modal attention method, multistream merger and classification technique are used to fuse multimodalities several times and extract key features; for the global-local session emotion feature extraction problem, graph construction, graph transformation network, graph convolutional network, and graph transformer are used, Bidirectional MOGLSTM and other strategies to effectively analyse the modal differences and conversational relationship differences existing between different modalities. The proposed method is analysed by IEMOCAP and MOSEI datasets, and the results show that the method improves the accuracy of multimodal session recognition, and effectively enhances the efficiency of session fusion recognition. The method in this paper does not focus on the parameter analysis of the algorithm, and the subsequent research will study how to use the parameter analysis to improve the model accuracy and efficiency.

Acknowledgements

This work was supported and funded by Scientific research project of Ningxia Education Department. (NO: NYG2024185)

6 REFERENCES

[1] Ni, J., Xie, W., Liu, Y., Zhang, J., Wan, Y., & Ge, H. (2024). Driver emotion recognition involving multimodal signals: Electrophysiological response, nasal-tip temperature, and vehicle behavior. *Journal of Transportation Engineering, Part A: Systems*, *150*(1). https://doi.org/10.1061/jtepbs.teeng-7802

[2] Cao, R., Li, L., Zhang, Y., Wu, J., & Zhao, X. (2023). A lightweight multimodal footprint recognition network based on progressive multi-granularity feature fusion. *International Journal of Pattern Recognition and Artificial Intelligence, 37*(14). https://doi.org/10.1142/s0218001423570124

[3] Nguyen, V. A. & Kong, S. G. (2023). Multimodal feature fusion for illumination-invariant recognition of abnormal human behaviors. *Information Fusion, 100*, 101949. https://doi.org/10.1016/j.inffus.2023.101949

[4] Islam, M. M., Nooruddin, S., Karray, F., & Muhammad, G. (2023). Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things. *Information Fusion, 94,* 17-31. https://doi.org/10.1016/j.inffus.2023.01.015

[5] Ding, H., Wang, S., Xie, Z., Li, M., & Ma, L. (2023). A fine-grained vision and language representation framework with graph-based fashion semantic knowledge. *Computers & Graphics, 10,* 115. https://doi.org/10.1016/j.cag.2023.07.025

[6] Tang, C., Qian, M., Jia, R., Liu, H., & Wang, B. (2022). Forearm multimodal recognition based on IAHP - entropy weight combination. *IET Biometrics, 12*(1), 52-63. https://doi.org/10.1049/bme2.12080

[7] Zhong, Z., Hou, Z., Liang, J., Lin, E., & Shi, H. (2023). Multimodal cooperative self - attention network for action recognition. *IET Image Processing, 17*(6), 1775-1783. Portico. https://doi.org/10.1049/ipr2.12754

[8] Wang, Y., Qiu, S., Li, D., Du, C., Lu, B.-L., & He, H. (2022). Multi-Modal Domain Adaptation Variational Autoencoder for EEG-Based Emotion Recognition. *IEEE/CAA Journal of Automatica Sinica, 9*(9), 1612-1626. https://doi.org/10.1109/jas.2022.105515

[9] Liu, Z., Cheng, Q., Song, C., & Cheng, J. (2023). Cross-scale cascade transformer for multimodal human action recognition. *Pattern Recognition Letters, 168*, 17-23. https://doi.org/10.1016/j.patrec.2023.02.024

[10] Hu, W. J., Yang, K. C., & Tan, Z. Y. (2023). Adaptive approximate nearest neighbour search algorithm based on k-means features. *Intelligent Computers and Applications, 13*(9), 80-84.

[11] Wu, X. & Wang, B. (2023). Simulation of fuzzy stochastic mining for big data based on plain Bayes. *Computer Simulation, 11*, 501-505.
[12] Wu, Y. H., He, Z., & Zhao, S. M. (2024). Research on classification method based on support vector machine and correlation imaging. *Advances in Lasers and Optoelectronics, 61*(10). https://doi.org/10.3788/LOP231483
[13] Hao, L., Zhiquan, F., & Qingbei, G. (2022). Multimodal cross-attention mechanism-based algorithm for elderly behavior monitoring and recognition. *Journal of Electronics, 34*, 1-13.
[14] Yin, G., Sun, S., Yu, D., Li, D., & Zhang, K. (2022). A Multimodal Framework for Large-Scale Emotion Recognition by Fusing Music and Electrodermal Activity Signals. *ACM Transactions on Multimedia Computing, Communications, and Applications, 18*(3), 1-23. https://doi.org/10.1145/3490686
[15] Sun, D., Li, Y., & Han, Y. (2023). A Reinforcement Learning-Based Multimodal Scenario Hazardous Behavior Recognition Method. *International Journal of Computational Intelligence Studies, 12*(2), 1. https://doi.org/10.1504/ijcistudies.2023.10053639
[16] Wen, J., Jiang, D., Tu, G., Liu, C., & Cambria, E. (2023). Dynamic interactive multiview memory network for emotion recognition in conversation. *Information Fusion, 91*, 123-133. https://doi.org/10.1016/j.inffus.2022.10.009
[17] Zhang, Y., Wang, J., Liu, Y., Rong, L., Zheng, Q., & Song, D. (2023). A multitask learning model for multimodal sarcasm, sentiment, and emotion recognition in conversations. *Information Fusion, 93,* 282-301. https://doi.org/10.1016/j.inffus.2023.01.005
[18] Yu, F., Wu, Y., Ma, S., Xu, M., Li, H., Qu, H., Song, C., Wang, T., Zhao, R., & Shi, L. (2023). Brain-inspired multimodal hybrid neural network for robot place recognition. *Science Robotics, 8*(78). https://doi.org/10.1126/scirobotics.abm6996
[19] Mukhtar, H. & Khan, M. U. G. (2024). CMOT: A cross-modality transformer for RGB-D fusion in person re-identification with online learning capabilities. *Knowledge-Based Systems, 283*. https://doi.org/10.1016/j.knosys.2023.111155
[20] Wang, B., Huang, X., Cao, G., Yang, L., Wei, X., & Tao, Z. (2022). Attention-enhanced and trusted multimodal learning for micro-video venue recognition. *Computers and Electrical Engineering, 102*. https://doi.org/10.1016/j.compeleceng.2022.108127
[21] Dubovi, I. (2022). Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology. *Computers & Education, 183*. https://doi.org/10.1016/j.compedu.2022.104495
[22] Becker, H. C., Lopez, M. F., King, C. E., & Griffin, W. C. (2023). Oxytocin Reduces Sensitized Stress-Induced Alcohol Relapse in a Model of Posttraumatic Stress Disorder and Alcohol Use Disorder Comorbidity. *Biological Psychiatry, 94*(3), 215-225. https://doi.org/10.1016/j.biopsych.2022.12.003
[23] Ding, J., Wang, Y., Si, H., Gao, S., & Xing, J. (2022). Multimodal Fusion-AdaBoost Based Activity Recognition for Smart Home on WiFi Platform. *IEEE Sensors Journal, 22*(5), 4661-4674. https://doi.org/10.1109/jsen.2022.3146137
[24] Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2024). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications, 237*. https://doi.org/10.1016/j.eswa.2023.121692
[25] Mocanu, B., Tapu, R., & Zaharia, T. (2023). Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and Vision Computing, 133*. https://doi.org/10.1016/j.imavis.2023.104676
[26] Li, J., Wang, X., Lv, G., & Zeng, Z. (2023). GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing, 550*. https://doi.org/10.1016/j.neucom.2023.126427
[27] Ezzameli, K. & Mahersia, H. (2023). Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion, 99*. https://doi.org/10.1016/j.inffus.2023.101847
[28] Devanga, S. R., Wilgenhof, R., & Mathew, M. (2021). Collaborative referencing using hand gestures in Wernicke's aphasia: Discourse analysis of a case study. *Aphasiology, 36*(9), 1072-1095. https://doi.org/10.1080/02687038.2021.1937919
[29] Xu, X., Li, X., Li, Q., & Yao, Z. (2023). Research on technology identification link prediction method based on graph neural network. *Data Analysis and Knowledge Discovery, 6*, 15-25.
[30] Fan, X. Y. & Zhang, H. R. (2023). Fault localization and fault type identification in distribution networks based on graph convolutional networks. *Experimental Technology and Management, 40*(1), 5.

Contact information:

Zhixue WANG
(Corresponding author)
School of Mathematics and Computer Science,
Ningxia Normal University,
Guyuan, 756000, Ningxia, China
E-mail: gu_wzx@163.com

Hongwu ZHANG
School of Mathematics and Computer Science,
Ningxia Normal University,
Guyuan, 756000, Ningxia, China
E-mail: 89299306@nxnu.edu.cn

Kai KANG
School of Mathematics and Computer Science,
Ningxia Normal University,
Guyuan, 756000, Ningxia, China
E-mail: kangkai@nxnu.edu.cn