ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

Multimodal conversational emotion recognition based on hier-archical Transformer

Bai Yimin[1,a], Zhang Pengwei[1,b], Zhang Jingze[2,c], Chen Jingxia[1,d,*]

[1]Shaanxi University of Science & Technology, Xi'an, China
[2]Xi'an Gaoxin No.1 Experimental High School, Xi'an, China
[a]221612100@sust.edu.cn, [b]zhangpengwei@sust.edu.cn, [c]169505153@qq.com,
[d]chenjingxia@sust.edu.cn
[*]Corresponding author

Abstract: *Addressing the issues of limited single-modality feature representation, inadequate multimodal feature fusion, and the difficulty of modeling conversation scenarios in multimodal conversational emotion recognition tasks, a Hierarchical Transformer-based multimodal conversational emotion recognition model has been proposed. The model employs self-attention bidirectional gated recurrent units to delve into the contextual dependencies of single-modality features such as text, video, and audio, thereby enhancing the representational power of features. Through hierarchical gated multi-head attention, it learns complementary information among modalities and adaptively learns the weights of each modality, reducing the noise interference of redundant information on multimodal features. The hierarchical Transformer is used to model conversation scenarios, utilizing a masking mechanism to simulate dependencies within contextual language, within speakers, and between speakers, gaining a deeper understanding of the emotional states of speakers in conversations. On the IEMOCAP and MELD benchmark datasets, the model achieved accuracy and F1 scores of 71.10% and 70.97% on IEMOCAP, and 67.16% and 66.11% on MELD, respectively, outperforming similar methods in terms of accuracy.*

Keywords: *Emotion Recognition in Conversation (ERC), Multimodality, Transformer, Gated Fusion, Multi-Head Attention*

1. Introduction

Emotions a Emotions are an integral part of human life and play a central role in daily communication, decision-making processes, and social interaction. In interpersonal interactions, emotions are expressed through various forms such as language, facial expressions, voice, and gestures. With the development of artificial intelligence technology, conversational emotion recognition (ERC) is gradually exerting great application value in the fields of opinion mining in social media [1] and empathic dialogue systems [2], especially in the field of human-computer dialogue interaction [3], which is getting more and more attention from researchers.

The research of ERC aims to model the relationship between the contextual context and the speaker in a conversation, so that the machine can understand and recognize the emotions expressed by humans during the communication process, thus providing a more intelligent, natural, and emotional interaction experience for hu-man-machine dialogues [4]. The dependencies that need to be modeled by ERC include self-dependence and environment dependence. Self-dependence means that the emotion of the current discourse is influenced by the emotion of the same speaker in the previous discourse. Since the speaker's affective states are continuous to some extent, their previous affective states need to be taken into account when analyzing the effect of the current discourse. Context dependency means that the emotion of the current discourse is influenced by other speakers or other factors in the conversational environment, such as the emotion of another speaker, the topic, and the overall conversational atmosphere.

Currently, ERC has received a lot of attention from researchers. Poria [5] et al. used Bi-LSTM to obtain contextual information on conversational sequences to enhance the understanding of context and emotion. Jiao et al [6] used two gated recurrent units (GRUs) to model the contextual relationships between words and discourse respectively. Hazarika et al [7-8] out CMN and ICON use GRUs to capture the current speaker's affective state and memory networks for storing and updating the previous dialogue history, which improves the emotion recognition by modeling the self-dependence and environment-dependence of a conversation, thus improving the emotion recognition accuracy. Majumder et al [9] proposed the DialogueRNN model uses three different GRUs to update the speaker, context, and sentiment states in the conversation respectively. Ghosal et al [10] proposed DialogueGCN model uses graph structure to consider speaker and dialogue sequence information for modeling. Li et al [11] proposed that the HiTrans Model consists of two layers of Transformer for obtaining global contextual information and modeling speaker-sensitive dependencies using an auxiliary task. Li et al [12] proposed TRMSM using textual modalities for sentiment recognition in three layers of Transformer modeling conversational context, intra-speaker, and inter-speaker. Hu J et al [13] proposed MMGCN model uses a graph convolutional network structure to model the speaker. Hu D et al [14] proposed MM-DFN to design a dynamic fusion module to fuse multimodal contextual information, reduce redundancy, and enhance inter-modal complementarity. Hu G et al [15] proposed the UniMSE model to model the speaker by fusing the modalities at the syntactic and semantic levels and using inter-modal contrast learning to differentiate sample representations through intra- and inter-modal interactions, inter-modal learning weight assignment, and enhancement of multimodal feature representations. Zhang et al [16] proposed the HAAN-ERC model to capture the intra- and inter-modal influences of the speaker and modality in each unimodal conversation using a hierarchical Transformer, and to fuse multimodal features through an adaptive attention mechanism. Du Jinming et al [17] proposed that the CK-ERC model uses a knowledge graph and a dynamic threshold-based course learning strategy to help the model accurately model spoken information. Feng Hongqi et al [18] effectively fused multimodal information by combining multilevel attention and multistream graph neural networks to capture global and local features of conversations. Liu Xinyu et al [19] used graph convolutional neural networks to construct graph structures for global conversations, speakers' own in-fluences, and inter-speaker influences to solve the problem of speaker personality modeling. Xudong Shen et al [20] proposed an MTDAG model using a temporal information-aware directed acyclic graph to capture multimodal rich feature information by optimizing discourse weights and fusing context and speaker information. In this paper, we propose a hierarchical Transformer to model the conversation scenario, using a masking mechanism to model the conversation contextual context, intra-speaker and inter-speaker.

Multimodal emotion recognition not only focuses on the emotional information of a single modality but also needs to process and fuse a large amount of information from different modalities. To reduce the noise interference caused by multimodal feature fusion, the complementary information between multimodalities is fully utilized. Early research on multimodal fusion mainly consists of integrating the features of different modalities at the input level constructing different models for each modality, and then integrating their outputs by methods such as majority voting or weighted averaging. These two methods are simple but cannot effectively capture in-tra- and inter-modal interactions. Model-level fusion became popular afterward. Zadeh et al [21] proposed the TFN model which uses tensor fusion networks to simulate the relationships between individual modalities, using high-dimensional tensors to represent multimodal features. MFN [22] learns cross-modal interactions through attentional mechanisms and stores information in controlled memories over time through multi-view gates. Liu et al [23] proposed the LMF model to reduce computational effort by performing a low rank through weights matrix decomposition to reduce computation. Tsai et al [24] modeled cross-modal remote dependencies using a cross-modal converter. Sahay et al [25] synthesized the advantages of both by using LMF to obtain multimodal features and fusing multimodal and unimodal interactions through a cross-modal attention mechanism. Rahman et al [26] designed a multimodal adaptive gate (MAG) that captures intra- and inter-modal interactions between conversational discourse while learning inter-modal weights. In this paper, we propose a hierarchical cross-modal multi-head attentional fusion method, which uses a hierarchical fusion approach to reduce the noise of the fused features and make full use of the complementary information of each modality.

In summary, multimodal conversational emotion recognition faces the challenge of how to fully fuse the feature information of different modalities, i.e., not only fully learning the complementary information between modalities but also removing the redundant information; and modeling the conversational scene to obtain rich emotional cues from the speaker. To address the above problems, this paper proposes a multimodal ERC model based on a hierarchical Transformer (HTMM-ERC). Firstly, the contextual relationships in the sequence of unimodal features are captured by SA-BiGRU, to better understand the complexity of emotional expressions, determine the importance of each modal data, and further reduce redundant information. Second, the intra- and inter-modal interrelationships of different discourse features are captured through a multi-head attention mechanism in the Transformer layer, and the inter-modal weights are adaptively learned using a gated network. Contextual interactions, internal dependencies, and interdependencies between speakers are modeled using a layered Transformer. Finally, the sentiment is categorized. The model in this paper is experimentally validated on two publicly available datasets, IEMOCAP [27] and MELD [28], demonstrating the effectiveness of this paper's

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

approach to the multimodal ERC task.

2. Related Work

We only accept papers written in English and without orthographic errors.

Please do not add any headers, footers and page numbers in the article, as we will do that uniformly.

All the text must use the font, Times New Roman. On Macintosh, please choose font, Times. Except in special circumstances, such as program code.

2.1. Task Definition

Denote the session dataset as $C = \{c_1, c_2, \ldots, c_P\}$, P denotes the number of sessions in the dataset. Each session $U = \{u_1, u_2, \ldots, u_N\}$, where N denotes the number of words in the session. Denote the speakers as $S = \{s_1, s_2, \ldots s_M\}$, where M denotes the number of speakers. Each discourse U_i has three modal features: textual (t), audio (a) and visual (v), denoted as $U_t = \{u_t^1, u_t^2, \ldots, u_t^N\}$, $U_a = \{u_a^1, u_a^2, \ldots, u_a^N\}$, $U_v = \{u_v^1, u_v^2, \ldots, u_v^N\}$, respectively. Each discourse corresponds to one speaker, and each speaker can correspond to multiple dis-courses. Each discourse has a corresponding sentiment label, denoted as $Y = \{y_1, y_2, \ldots, y_N\}$, and the goal of the ERC task is to predict the sentiment label y_i' of each discourse u_i.

2.2. Model overview

The HTMM-ERC model proposed in this paper has four main parts: (1) Modal encoder module, which firstly obtains the unimodal feature vectors from the raw data through preprocessing, and then uses the SA-BiGRU module to capture the contextual relationships in the feature sequences and determine the importance of each modal data, to understand the complexity of the emotional expression better and further reduce the redundant information. (2) A layered cross-modal gated fusion module that captures the intra- and inter-modal interrelationships of different discourses through a multi-head attention mechanism in the Transformer layer, and then adaptively learns the inter-modal weights through the gated network. (3) Hierarchical Transformer conversation modeling module, which models global dependencies, speaker's dependencies, and inter-speaker dependencies through a hierarchical Transformer for a comprehensive and in-depth understanding of the conversational context. (4) Prediction and Classification, using classifiers for 6 categories of sentiment classification. The general framework diagram of the HTMM-ERC model is shown in Figure 1.

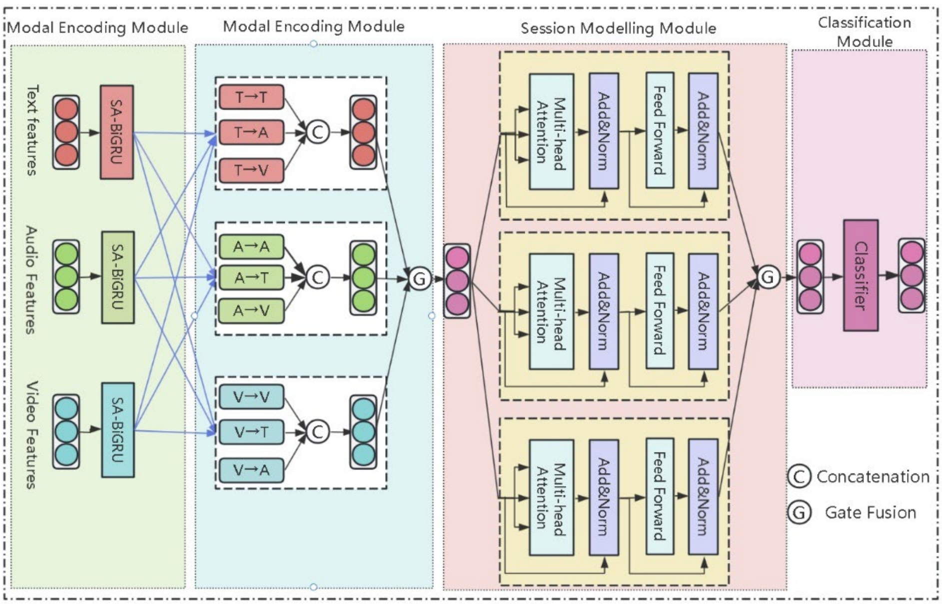

Figure 1: Framework diagram of the HTMM-ERC model.

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

2.3. Model overview

There are interdependencies between each discourse of a session, SA-BiGRU (Self-Attention-Based Bidirectional Gated Recurrent Unit) is a module that combines a Bidirectional Gated Recurrent Unit (BiGRU) and a self-attention mechanism. This module achieves an effective capture of contextual relationships in feature sequences by implementing them in BiGRU. The self-attention mechanism is introduced at the output layer to further reduce the redundant information, which helps to better understand the complexity of sentiment expression and improve the accuracy and reliability of sentiment analysis. The SA-BiGRU network structure is shown in Figure 2.

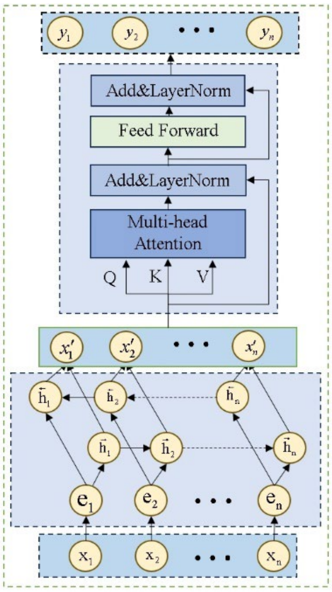

Figure 2: Network structure diagram of SA-BiGRU.

BiGRU is a simplified version of BiLSTM, which has a simple structure containing only reset and update gates, reducing network complexity while improving computational efficiency. It can capture the long-term dependencies in the sequence and better the temporal characteristics of the data. The network structure expression of BiGRU is as follows:

$$\vec{h}_{\mathrm{t}} = GRU\left(x_t, \vec{h}_{t-1}\right) \tag{1}$$

$$\overleftarrow{h}_{\mathrm{t}} = GRU\left(x_t, \overleftarrow{h}_{t-1}\right) \tag{2}$$

$$h_{\mathrm{t}} = W_{\vec{h}_t}\vec{h}_t + W_{\overleftarrow{h}_t}\overleftarrow{h}_t + b_t \tag{3}$$

where: $\mathrm{W}_{\vec{h}_t}, W_{\overleftarrow{h}_t}, \vec{h}_t, \overleftarrow{h}_t$ denotes the forward and backward hidden layer states and weights at moment t, respectively, and b_t denotes the bias of the hidden layer state at moment *t*.

The self-attention mechanism captures the key information between positions in a unimodal sequence of features. To improve the ability to mine the deeper features of the data, for the input feature vectors, their corresponding generation vectors Q, K, and V, are calculated as follows:

$$Attention\left(Q, K, V\right) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)\cdot V \tag{4}$$

Where: d_k denotes the dimension of the key vector.

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

2.4. Cross-modal feature fusion module based on hierarchical Transformer

Unimodal features cannot adequately express complex emotional information, different modalities can provide complementary information, text can provide semantic information, while images and speech can provide emotional, intonational or visual information. According to Rahman et al [26], this paper designs a hierarchical cross-modal interaction and gating fusion module that includes both unimodal and multimodal modalities, which enables full interaction of multimodal information and reduces the noise and redundant information of unimodal features.

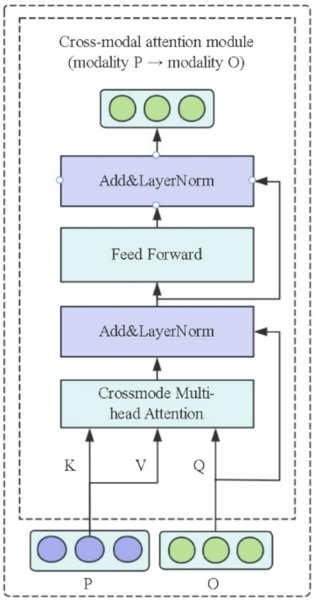

Figure 3: Structure Diagram of Cross-modal Fusion Module.

As shown in Figure. 3 is the schematic structure of the cross-modal fusion module, which uses the Transformer encoder [29] for unimodal and cross-modal interactions, which contains three input query vectors *Q*, key vectors *K*, and value vectors *V*, denoted $TransformerEncode(Q,K,V)$, simplified $TE(Q,K,V)$, and for unimodal can be expressed as :

$$Z_{m\to m}=TE(U_m,U_m,U_m) \tag{5}$$

where: U_m is the $m\in(t,a,v)$ modal discourse feature representation. The cross-modal representation of *m* to *n* different modalities is:

$$Z_{m\to n}=TE(U_m,U_n,U_n) \tag{6}$$

The six sets of modal interaction feature of text-to-audio and video, audio-to-text and video, and video-to-text and audio are obtained by cross-modal attention as follows:

$$Z_{t\to a}=TE(U_t,U_a,U_a) \tag{7}$$

$$Z_{t\to v}=TE(U_t,U_v,U_a) \tag{8}$$

$$Z_{a\to t}=TE(U_a,U_t,U_t) \tag{9}$$

$$Z_{a\to v}=TE(U_a,U_v,U_v) \tag{10}$$

$$Z_{v\to t}=TE(U_v,U_t,U_t) \tag{11}$$

$$Z_{v\to a}=TE(U_v,U_a,U_a) \tag{12}$$

In order to obtain a complete representation of the unimodal state, the three unimodal and six cross-modal interaction feature matrices are spliced and represented as:

$$Z_T=[Z_{t\to t}\oplus Z_{t\to a}\oplus Z_{t\to v}] \tag{13}$$

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

$$Z_A = [Z_{a \to a} \oplus Z_{a \to t} \oplus Z_{a \to v}] \tag{14}$$

$$Z_V = [Z_{v \to v} \oplus Z_{v \to t} \oplus Z_{v \to a}] \tag{15}$$

Where: $[\oplus]$ denotes the splicing operation.

The gating mechanism is a commonly used method to control the flow of information between different parts of a network. In this paper, we use the SoftMax function to implement a dynamic gating mechanism to adapt weight learning in multimodal emotion recognition tasks. The specific steps are to map the output vectors of each modality into scalar values through the fully connected layer and input the SoftMax function to get the weight distribution, which is used to weigh the output vectors and form the weighted sum as the final result. Gated fusion can dynamically learn the weights between each modality to handle multimodal data better. The final multimodal features of the discourse are calculated as follows:

$$r_i = W_i * Z_i + b_i \tag{16}$$

$$g_i = Soft\max(r_i) \tag{17}$$

$$Z_i^{'} = \sum_{i \in \{t,a,v\}} g_i \otimes Z_i \tag{18}$$

Where: W_i and b_i are the weight matrix and bias term respectively. i denotes as different modalities.

2.5. Cross-modal feature fusion module based on hierarchical Transformer

The core of the ERC task is to model the contextual context and the relationship between speakers in a session, and with the success of Transformer [30] in multiple domains, the attention mechanism in Transformer cleverly solves the problem of modelling context in a session. According to Zhang et al [16], this paper argues that the fused multimodal features have more comprehensive and rich emotional information, so the multimodal features should be modelled as a session scenario. According to the pair of research ideas of Li et al [12], this paper proposes a session modelling module based on a layered Transformer, which cleverly uses the Transformer mask mechanism to construct three different masks to simulate global dependencies, speaker's dependencies and inter-speaker dependencies. That is, there are three Transformer encoding layers: the first layer uses global masks to model the session context and capture the global dependencies of the session. The second layer models the speaker itself using a speaker self-mask, which is used to capture intra-speaker dependencies. The third layer models the inter-speaker using different inter-speaker masks for capturing and analyzing interactions, sentiment transfer and dependencies between different speakers. The three coding layers produce three features modelled using different conversational scenarios. In this paper, we use the same gating mechanism as multimodal feature fusion to adaptively and dynamically learn the weights of each feature, and ultimately obtain a multimodal feature vector containing rich conversational information. The specific calculation method is shown in Eq. (16) to Eq. (18).

The following are the main components of the Transformer coding layer:

1) Positional encoding: it can encode the relative positional relationship of input features so that the model can clarify the contextual relationship and thus better understand the semantics. The calculation formula is as follows:

$$\begin{aligned} PE_{(pos,2i)} &= \sin(\frac{pos}{10000^{2i/d}}), \\ PE_{(pos,2i+1)} &= \cos(\frac{pos}{10000^{2i/d}}), \end{aligned} \tag{19}$$

2) Multi-head attention: it consists of multiple self-attention modules, which can process the target discourse in parallel to learn a more comprehensive feature representation, and project the discourse vectors to query, key and value representations respectively, and the self-attention with masking is calculated as follows:

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

$$A(Q,K,V,M) = Soft\max(\frac{(QK^T)*M}{\sqrt{d_k}}) \cdot V \tag{20}$$

Where: $Q \in \mathbb{R}^{N \times d_k}, \mathrm{K} \in \mathbb{R}^{N \times d_k}, \mathrm{V} \in \mathbb{R}^{N \times d_k}$ denotes query, key and value respectively, d_k denotes the dimension of key vector, $[*]$ denotes element multiplication, $[\cdot]$ denotes dot product, $\mathrm{M} \in \mathbb{R}^{\mathrm{N} \times \mathrm{N}}$ denotes mask matrix.

3) Normalisation layer, feed-forward fully connected layer: the normalisation layer is used to normalise the output of each sub-layer, which helps to accelerate training and improve the generalization ability of the model. The feed-forward fully connected layer introduces nonlinearity so that the model can better fit complex input-output relationships. The computation is performed as:

$$A = LN(U + X) \tag{21}$$

$$F = \mathrm{Re}\,lu(AW_1)W_2 \tag{22}$$

$$F' = LN(A + F) \tag{23}$$

Where *LN* is the layer normalisation, *Relu* is the activation function, W_1 and W_2 are the weight matrices of the feed-forward fully connected layer, *U* denotes the input features, and *X* denotes the output of the multi-head attention module.

The three mask-creation methods used in this paper are as follows:

1) Global Mask: This mask sets all own elements to True, which means that each target discourse has access to (i.e., can see) all contextual discourse in a multi-head attention computation. This mask allows the model to take into account the entire context of the dialogue when processing each target discourse, leading to a better understanding of the meaning of the target discourse.

2) Speaker's own mask: this mask sets the value to True if the speaker of the discourse has the same speaker label as the target discourse, and False if the speaker of the discourse does not have the same speaker label as the target discourse. This mask handles the coherence of the expression of the same speaker in different discourses and ensures that the model focuses only on what is relevant to that speaker when processing the target discourse.

3) Inter-speaker mask: this mask sets the value of the position where the speaker of the discourse has the same speaker label as the speaker of the target discourse to False and sets the other positions to True. This mask encourages the model to focus on other speakers related to the target speaker, and the inter-speaker mask helps the model to understand the interactions and emotion transfer among different speakers.

2.6. Module for categorising emotions

In order to get the probability of the sentiment category, in this paper, the (gated fusion layer processed feature vector) is input into a classifier with fully connected (FC) layer and SoftMax layer for prediction:

$$P'_i = \mathrm{Re}\,lu(W_H H_i + b_H) \tag{24}$$

$$P_i = SoftM\,\mathrm{ax}(W_P P'_i + b_P) \tag{25}$$

$$\hat{\mathrm{y}}_i = \arg\max_{\mathrm{k} \in Y}(P_i[k]) \tag{26}$$

At training time, this paper uses cross-entropy loss to measure the quality of sentiment prediction:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j} \log(\hat{y}_{i,j}) \tag{27}$$

Where: N is the number of dialogues, C is the number of words in dialogue *i*, $y_{i,j}$ is the true sentiment label of word *j* in dialogue *i*, and $\hat{y}_{i,j}$ denotes the probability distribution of the predicted sentiment label of word *j* in dialogue *i*. The Adam optimiser should be employed for the purpose of

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

training the network model.

3. Datasets and data pre-processing

3.1. Datasets

We hope you find the information in this template in this paper, we validate and evaluate the performance of the proposed method on two publicly available datasets: the IEMOCAP dataset [27] and the MELD dataset [28], both of which are multimodal conversational emotion recognition datasets containing three modalities. The statistical results of the two datasets are shown in Table 1.

IEMOCAP dataset: consists of 10 binary dialogues performed by actors according to a script, containing 153 dialogues and 7433 words. IEMOCAP is divided into 5 sessions, where the first 4 sessions are used for training and the last one is used for testing. Each sentence in the dialogue was tagged with one of the six category emotion la-bels, namely happy, sad, neutral, angry, excited and frustrated.

MELD dataset: this is a multi-speaker multimodal pair conversation dataset collected from the Friends TV series, which contains 1433 dialogues and 13708 utterances. Each sentence is labelled with one of seven emotions: Neutral, Surprise, Fear, Sadness, Joy, Disgust and Anger.

Table 1: Statistical Information of IEMOCAP and MELD.

Dataset	**Conversations**			**Utterances**			**Classes**
	Train	Val	Test	Train	Val	Test	
IEMOCAP	120		31	5810		1623	6
MELD	1039	114	280	9989	1109	2610	7

3.2. Data pre-processing

3.2.1. Text feature extraction

Text modal features are extracted using the large-scale pre-trained model RoBERT-Large [31] based on Transformer. RoBERT model is based on the improvement of the BERT model, which uses more data, and the larger batch makes it possible to mine the text data for syntactic and sentence-level features, thus achieving a more powerful characterisation capability, and finally obtains 1024-dimensional sentence-level features.

3.2.2. Speech Feature Extraction

Speech modal features are configured by Hazarika [4] et al. Frame level speech features are extracted using OpenSmile based on IS13 profiles and then sentence level features are obtained through the fully connected layer. The dimensionality of the speech features on the IEMOCAP dataset is 1582 and the dimensionality of the speech features on the MELD dataset is 300.

3.2.3. Image Feature Extraction

Image modal features are pre-trained on facial expression recognition corpus using DenseNet [32] and then extracted on the ERC dataset DenseNet is a CNN network that achieves feature reuse and optimised information flow through densely connected layers with a small number of parameters, superior performance and compactness, the final dimensionality of the facial expression features of the output image is 342.

4. Results

4.1. Study parameter settings

In this experiment, AdamW is chosen as the optimiser in this paper to achieve more stable convergence during the training process. For the IEMOCAP dataset, the batch size (batch size) is set to 16 and the initial learning rate is set to 1E-3. For the MELD dataset, this paper sets the batch size to 32 and adjusts the initial learning rate to 1E-4. The whole training process is iterated 50 times, and at the same time, the Dropout is set to 0.5 to prevent overfitting. In addition, 8 heads were used by default to implement the multi-head attention mechanism. All experiments were performed on an NVIDIA GeForce RTX 3090 GPU equipped with 24 GB of video memory. In this paper, Accuracy (Accuracy, Acc) and F1

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

value (F1-score) are used as evaluation indexes.

4.2. Study results

A graph of the loss trend during training on IEMOCAP and MELD is shown in Fig. 4. There is a trend of a gradual decrease in the model loss values with the increase in the number of training rounds. At the beginning of the training period, the loss values are high, but as the training progresses, the loss values gradually decrease, indicating that the model is converging gradually. As shown in Table 2, the HTMM-ERC model achieves an average accuracy and average weighted F1 score of 71.10% and 70.97% on the IEMOCAP dataset, and an average accuracy and average weighted F1 score of 67.16% and 66.11% on the MELD dataset.

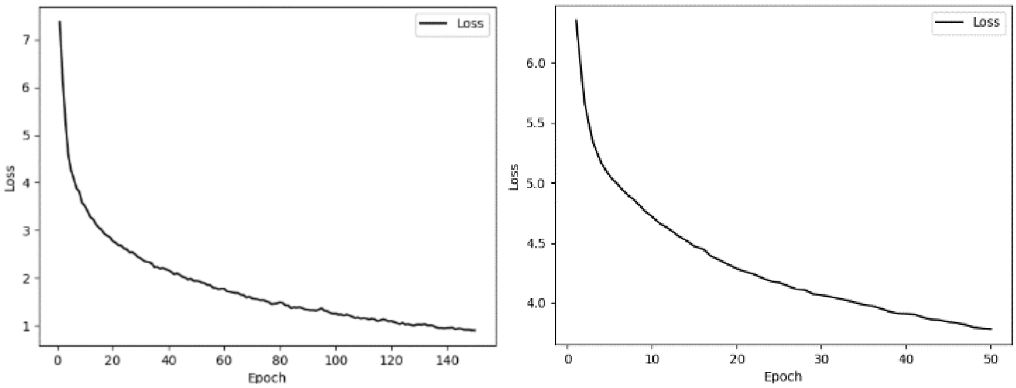

Figure 4: The loss curve during training on IEMOCAP and MELD: (a) Training loss for the IEMOCAP dataset sheets; (b) Training loss for the MELD dataset sheets.

4.3. Contrastive Models

1) DialogueRNN [9]: a Recurrent Neural Network (RNN) based method for modelling context and speaker information. The method uses independent gated recurrent units (GRUs) to model the speaker state and contextual information etc. respectively.

2) DialogueGCN [10]: Graph Convolutional Networks (GCNs) are applied to the task of emotion recognition by extracting discourse-level features using Bidirectional Long Short-Term Memory Networks (Bi-LSTMs) and constructing graph data based on these features. The method portrays the conversational context of emotion recognition by modelling the self-dependence of the interlocutor and the dependencies between speakers.

3) MMGCN [13]: this method constructs a conversational graph based on all three modalities and designs a multimodal fusion graph convolutional network to model contextual dependencies across multiple modalities.

4) DialogueTRM [33]: a hierarchical Transformer model is used to deal with different context preferences in each modality, and a multi-granularity interactive fusion strategy is designed to learn the contributions of discourse across modalities.

5) MM-DFN [14]: a graph-based dynamic fusion module is designed to fuse multimodal contextual features to reduce redundancy and enhance inter-modal complementarity.

6) UniMSE [15]: fuses acoustic and visual modal features with multilevel textual features using the T5 model for inter-modal comparison learning to obtain a more discriminative multimodal representation.

7) HAAN-ERC [16]: employs a hierarchical Transformer to capture speaker and modal internal interactions in individual unimodal dialogue contexts and fuses them through an adaptive attention mechanism.

8) MTDAG [20]: a directed acyclic graph model for temporal information perception proposed in this study by optimising discourse weights, fusing context and speaker information as well as capturing multimodal information effectively.

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

4.4. Comparison Study Results

The HTMM-ERC model proposed in this paper addresses the challenges of the difficulty of fusing multimodal data for multimodal session emotion recognition and the complexity of modelling session scenarios, and in order to further validate the excellence of the model, Table 2 demonstrates the results of the comparative experiments between the HTMM-ERC model proposed in this study and the other baseline models on both the IEMOCAP and MELD datasets. As can be seen from the table, the HTMM-ERC model outperforms the other models on both datasets.

On the IEMOCAP dataset, the HTMM-ERC model achieves an average accuracy (ACC) of 71.10%, which is improved by 3.37%, 2.78%, and 4.49% compared to the DialogueRNN, DialogueGCN, and MMGCN models, respectively. Meanwhile, the average weighted F1 score (w-F1) of the HTMM-ERC model also reaches 70.97%, which is improved by 3.20%, 2.59% and 4.76% compared to the above models, respectively.

On the MELD dataset, the HTMM-ERC model has an average accuracy (ACC) of 67.16%, which is an improvement of 1.24%, 1.14%, and 5.82% compared to the DialogueRNN, DialogueGCN, and MMGCN models, respectively. In addition, the average weighted F1 score (w-F1) of the HTMM-ERC model is 66.11%, which is improved by 1.04%, 1.10%, and 7.67% compared to the above models, respectively.

Table 2: Comparison of Experimental Results on IEMOCAP and MELD Datasets.

Models	**IEMOCAP(Avg)**		**IEMD(Avg)**	
	ACC	w-F1	ACC	w-F1
DialogueRNN[9]	67.73	67.77	65.92	65.07
DialogueGCN[10]	68.32	68.38	66.02	65.01
MMGCN [13]	66.61	66.25	61.34	58.41
DialogueTRM[33]	68.52	68.20	65.10	63.80
MM-DFN[14]	68.21	68.18	62.49	59.46
UniMSE[15]	70.56	70.66	65.09	65.51
HAAN-ERC[16]	69.48	69.47	66.31	65.50
MTDAG[20]	70.98	70.75	65.36	64.51
HTMM-ERC (ours)	**71.10**	**70.97**	**67.16**	**66.11**

In summary, the HTMM-ERC model proposed in this study outperforms other baseline models for emotion recognition on both IEMOCAP and MELD datasets, and this model significantly reduces redundant information by obtaining the contextual relationships of unimodal feature vectors, effectively focusing on key features in emotion recognition. It makes use of the Transformer layer's multi-head attention mechanism and gating network to adaptively learn modal weights, comprehensively capture intra- and inter-modal relationships, and deepen conversational contextual understanding by modelling conversational global as well as speaker interdependencies through the hierarchical Transformer encoder, and proving its effectiveness in processing temporal, contextual and multimodal information.

4.5. Ablation Study

In order to investigate the effect of different modules and different modalities in the HTMM model, ablation experiments were conducted on both datasets, considering the following settings:

1) w/o ConM: remove the global mask in the session modelling module used.

2) w/o SpIntra: removing the speaker's own mask in the session modelling module used.

3) w/o SpInter: removes the inter-speaker mask in the session modelling module used.

4) w/o ConEn: removes the entire session modelling module.

5) w/o CRFU: remove the layered cross-modal fusion module.

Table 3 shows the results of the ablation experiments for the HTMM-ERC model on two datasets. The experimental results reveal that, in the context of the session modelling module, the model's performance undergoes a decline when the global mask, the speaker's mask, and the inter-speaker mask are removed individually. More notably, the most significant performance drop occurs when the entire session modelling module is removed, highlighting the indispensable and effective nature of all its components within the HTMM-ERC model. Furthermore, the HTMM-ERC model's components are

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

indispensable and effective. When the hierarchical cross-modal fusion module is excluded, and multimodal features are fused directly through concatenation, the model's performance decreases substantially, particularly on the MELD dataset. This underscores the crucial role played by the cross-modal fusion module in enhancing model performance. Additionally, using speech modality and image modality in isolation results in a significant decline in the model's accuracy and weighted F1 value, with the effect of image modality being particularly pronounced. In contrast, the performance degradation is relatively minor when text modality is used alone, suggesting that text modality exhibits greater stability in emotion recognition. Overall, the combination of the cross-modal fusion module and multimodality plays a pivotal role in the HTMM-ERC model's performance in emotion recognition tasks.

Table 3: Ablation Experimental Results of HTMM-ERC on Two Datasets.

	IEMOCAP(Avg)		**MELD(Avg)**	
	ACC	**w-F1**	**ACC**	**w-F1**
HTMM-ERC	**71.10**	**70.97**	**67.16**	**66.11**
w/o ConM	70.64	69.88	66.86	65.43
w/o SpIntra	69.71	70.20	66.70	65.53
w/o SpInter	70.15	69.84	67.09	65.95
w/o ConEn	69.15	69.02	66.42	65.25
w/o CRFU	69.30	69.43	65.97	65.42
A&T	70.42	70.11	66.85	65.90
A&V	60.78	60.98	45.69	42.18
T&V	69.05	69.13	66.77	65.52
T	68.69	68.54	66.52	65.83
A	59.67	59.48	50.45	50.06
V	40.52	41.06	37.72	33.58

5. Conclusions

The HTMM-ERC model proposed in this paper effectively captures complex emotional information in multimodal data while mitigating the interference of redundant information, through the synergistic action of three core modules: modal encoder, hierarchical cross-modal attention fusion, and session scene modeling. Validated through comparative and ablation experiments, the model demonstrates significant superiority in terms of performance and robustness. However, there is still room for improvement in this study, such as limitations in the dataset and high model complexity. In the future, we will explore new datasets, optimize the model structure to reduce computational costs, and integrate other advanced methods, such as knowledge graphs and reinforcement learning, to further improve the accuracy and robustness of emotion recognition.

References

[1] Chatterjee A Narahari K N, Joshi M, et al. SemEval-2019 task 3: EmoContext contextual emotion detection in text[C]//Proceedings of the 13th international workshop on semantic evaluation. 2019: 39-48.

[2] Zhou L, Gao J, Li D, et al. The Design and Implementation of XiaoIce, an Empathetic Social Chat-bot.2018[2024-03-27]. DOI: 10.48550/arXiv. 1812. 08989.

[3] Huang Zhong. Research on Expression Recognition and Reproduction Methods for Humanoid Robots[D]. Hefei University of Technology, 2017.

[4] Hazarika D, Poria S, Zadeh A, et al. Conversational memory network for emotion recognition in dyadic dialogue videos[C]//Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting. NIH Public Access, 2018, 2018: 2122.

[5] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers). 2017: 873-883.

[6] Jiao W, Yang H, King I, et al. HiGRU: Hierarchical Gated Recurrent Units for Utterance-Level Emotion Recognition[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 397-406.

[7] Hazarika D, Poria S, Zadeh A, et al. Conversational memory network for emotion recognition in

dyadic dialogue videos[C]//Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting. NIH Public Access, 2018, 2018: 2122.

[8] Hazarika D, Poria S, Mihalcea R, et al. Icon: Interactive conversational memory network for multimodal emotion detection[C]//Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 2594-2604.

[9] Majumder N, Poria S, Hazarika D, et al. Dialoguernn: An attentive rnn for emotion detection in conversations[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 6818-6825.

[10] Ghosal D, Majumder N, Poria S, et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, 2019.

[11] Li J, Ji D, Li F, et al. HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Con-versations[C]// Proceedings of the 28th International Conference on Computational Linguis-tics.2020.DOI: 10.18653/v1/2020.coling-main.370.

*[12] Li J, Lin Z, Fu P, et al. A hierarchical transformer with speaker modeling for emotion recognition in conversation[J]. ar**v preprint ar**v: 2012.14781, 2020.*

[13] Hu J, Liu Y, Zhao J, et al. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation[C] //Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 5666-5675.

[14] Hu D, Hou X, Wei L, et al. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 7037-7041.

[15] Hu G, Lin T E, Zhao Y, et al. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 7837-7851.

[16] Zhang T, Tan Z, Wu X. HAAN-ERC: hierarchical adaptive attention network for multimodal emotion recognition in con-versation[J]. Neural Computing and Applications, 2023, 35(24): 17619-17632.

[17] Du Jinming, Sun Yuanyuan, Lin Hongfei, et al. Integrating Knowledge Graphs and Curriculum Learning for Conversational Emotion Recognition[J]. Journal of Computer Research and Development, 2024, 61(05): 1299-1309.

[18] Feng Hongqi, Guo Yongxiang, Zhang Denghui, et al. Multimodal Conversation Emotion Recognition Combining Multi-Level Attention and Multi-Stream Graph Neural Networks[J/OL]. Computer Engineering and Applications: 1-11 [2024-05-16].http://kns.cnki.net/kcms/ detail/11.2127.TP. 20231218. 1335.014.html.

[19] Liu Xinyu, Xia Hongbin, Liu Yuan. A Conversational Emotion Recognition Model with Speaker Feature Fusion[J/OL]. Journal of Miniature Microcomputer Systems: 1-8 [2024-03-27]. http://kns.cnki. net/kcms/detail/21.1106.TP.20240229.1556.002.html.

[20] Shen Xudong, Huang Xianying, Zou Shihao. A Multimodal Conversational Emotion Recognition Model Based on Tempo-rally Aware DAG[J]. Journal of Computer Applications Research, 2024, 41(01): 51-58.

[21] Zadeh A, Chen M, Poria S, et al. Tensor Fusion Network for Multimodal Sentiment Analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.2017: 1103-1114.

[22] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).

[23] Liu Z, Shen Y, Lakshminarasimhan V B, et al. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2247-2256.

[24] Tsai Y H H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the conference. Association for computational linguistics. Meeting. NIH Public Access, 2019, 2019: 6558.

[25] Sahay S, Okur E, Kumar S H, et al. Low Rank Fusion based Transformers for Multimodal Sequences[C]//Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML). 2020: 29-34.

[26] Rahman W, Hasan M K, Lee S, et al. Integrating multimodal information in large pretrained transformers[C]//Proceedings of the conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2020, 2020: 2359.

ISSN 2616-5775 Vol. 8, Issue 1: 19-31, DOI: 10.25236/AJCIS.2025.080104

[27] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database [J]. Language resources and evaluation, 2008, 42: 335-359.

[28] Poria S, Hazarika D, Majumder N, et al. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 527-536.

[29] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.Linguistics.2020.

[30] Kalyan K S, Rajasekharan A, Sangeetha S. AMMU: a survey of transformer-based biomedical pretrained language models[J]. Journal of biomedical informatics, 2022, 126: 103982.

[31] Zhu Y, Xu Y, Yu F, et al. Deep Graph Contrastive Representation Learning[J]. 2020.DOI:10. 48550/arXiv.2006.04131.

[32] Huang G, Liu Z, Laurens V D M, et al. Densely Connected Convolutional Networks[J]. IEEE Computer Society, 2016.

[33] Mao Y, Liu G, Wang X, et al. DialogueTRM: Exploring Multi-Modal Emotional Dynamics in a Conversation[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 2694-2704.