IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received ***, accepted ***, date of publication ***, date of current version ***.

Digital Object Identifier

MM DialogueGAT- A Fusion Graph Attention Network for Emotion Recognition using Multi-model System

Rui Fu[1], Xiaomei Gai[2], Ahmed Abdulhakim Al-Absi [3], Mohammed Abdulhakim Al-Absi[4], Muhammed Alam[5],Ye Li[2],Meng Jiang[1], Xuewei Wang[2]

[1]College of Language Intelligence at Sichuan International Studies University, Language & Brain Research Center, Sichuan International Studies University, Chongqing, 400031, China;
[2]Shandong Province University Laboratory for Protected Horticulture, Blockchain Laboratory of Agriculture and Vegetables,Weifang University of Science and Technology,Weifang, 262700, China;
[3]Department of Smart Computing, Kyungdong University, 46 4-gil, Bongpo, Gosung, Gangwon-do 24764, Korea
[4]Department of Computer Engineering, Graduate School, Dongseo University, 47 Juryero, Sasanggu, Busan, 47011, Korea
[5]School of Engineering, London South Bank University, UK

Corresponding author: X.W.W,M.J (e-mail: wangxuewei@wfust.edu.cn. jiangmeng1973@163.com).

This research was funded by Shandong Provincial Natural Science Foundation(ZR20220920030), Shandong Provincial Natural Science Foundation(ZR2023MF048), Ministry of Education Science and Technology Development Center Innovation Fund(2022BL023), Weifang Soft Science(2022RKX108), Shandong Federation of Social (2023JCXK016), Weifang Uiversity of Science and Technology Doctoral Fund Project (2021KJBS13) , Weifang Uiversity of Science and Technology Doctoral Fund Project (2021RWBS07)

ABSTRACT Emotion recognition is an important part of human-computer interaction and human communication information is multi-model. Despite advancements in emotion recognition models, certain challenges persist. The first problem pertains to the predominant focus in existing research on mining the interaction information between modes and the context information in the dialogue process but neglects to mine the role information between multi-model states and context information in the dialogue process. The second problem is in the context information of the dialogue where the information is not completely transmitted in a temporal structure. Aiming at these two problems, we propose a multi-model fusion dialogue graph attention network (MM DialogueGAT). To solve the problem 1, the bidirectional GRU mechanism is used to extract the information from each model. In the multi-model information fusion problem, different model configurations and different combinations use the cross-model multi-head attention mechanism to establish a multi-head attention layer. Text, video and audio information are used as the main and auxiliary modes for information fusion. To solve the problem 2, in the temporal context information extraction problem, the GAT graph structure is used to capture the context information in the mode. The results show that our model achieves good results using the IMEOCAP datasets.

INDEX TERMS Emotion Recognition, Interaction Information, Multi-head Attention, GAT, Graph Attention Network.

I. INTRODUCTION

With the development of natural language processing NLP technology, people gradually realize that recognizing emotions is essential to better understand and process natural language. As the number of publicly available conversation data on social media platforms such as Facebook, Twitter, Youtube, and Reddit increased, so did the emphasis on emotion recognition in conversations.

In the early research work on emotion recognition ERC, the recurrent neural network RNN modeled the context information and sequential information to determine the emotion type according to the text content. Chung fed the dialogue information back into the two-way gating unit (GRU) for information modeling. Theoretically, recurrent neural network RNN, long short-term memory LSTM and bidirectional gating unit GRU can propagate contextual information, but only considering the sequential propagation of information affects the accuracy of recurrent neural network models in sentiment recognition and classification tasks.

With the development of neural networks, context information can not only be propagated sequentially, but also jumped and connected by information. DialogueRNN uses two GRUs to track each speaker's context information and global state as separate entities. ConGCN uses graph constitutional networks on both audio and text discourse features to model both speaker-discourse and discourse-discourse relationships in a single network. DialogueGCN treats a conversation as a graph structure where each statement or conversation wheel is represented as a node and the edges between nodes represent

their semantic or contextual relationships. M2FNet is a multi-modal fusion network for extracting emotional features from audio, video and text. However, most of the existing research on emotion recognition focuses on mining the interaction information between modes and the context information in the dialogue process and ignores the correlation information between the role information between multiple models and the context in the dialogue process. Secondly, in the context information of the dialogue, the information is not completely transmitted in a temporal structure and there will be some jumping connections of information.

In order to solve the **problem 1** of correlation information between multi-modal information and context in the dialogue process, we combine different modal information using deep learning methods to enable interactive communication between various modalities. The way a multi-head attention layer is introduced in the multi-modal fusion approach to fusion multi-modal characteristics of media content (text, audio, video).

In order to solve the **problem 2** of jump connection in the context information of the dialogue, we should combine the local and global information together to understand the emotional content of the dialogue more comprehensively and perform emotion recognition more accurately. Local information constituted of internal dependence and external dependence. The internal dependence means one speaker in the dialogue may affect the emotional changes of other speakers, the external dependence means the speaker may be affected by external factors. Global information in context refers to the combination of temporal contextual information in the conversation which contains the overall sentiment and development trend of the dialogue. All the local information is combined together with global information. This paper proposes a multi-modal method DialogueGAT, which uses the attention mechanism to calculate the attention weight of each node and its neighbor node and then weighted aggregation of node features and attention weight to improve the expressive ability and stability of the model.

Through the demonstration of comparative experiments and ablation experimental studies, the proposed MM DialogueGAT model achieved good experimental results on the multi-modal emotion recognition datasets IEMOCAP. Here are our main contributions:

- A multi-head attention layer is established based on different modal configurations and various combinations, achieving the capture of contextual information and multi-modal interaction between modes.
- A multi-modal fusion network called MM DialogueGAT is proposed for multi-modal in-depth interaction, GAT introduces an attention mechanism to dynamically weight the importance of node neighbors, so that nodes pay more attention to neighbor nodes when updating and use graph neural networks to capture the jump information in dialogue information, which strengthens the local information and global information. The complex dependency in dialogue is solved for the emotion recognition problem in dialogue. Experiments are carried out on multiple datasets and the superiority of the model is proved.

We explains the content according to the framework of the following chapters: the first section summarizes the paper. The second section discusses the work related to emotion recognition under single-modal and multi-modal conditions. The third section elaborates on the specific algorithm. Section 4 describes the experimental setup. Section 5 presents and explains the experimental results. Section 6 summarizes the papers.

II. RELATED WORK

A. UNI-MODAL EMOTION RECOGNITION

The single-modal approach tends to take text as the main mode of communication and only uses a single modality to complete the emotion recognition task, while multi-modal emotion recognition refers to the task of understanding and inferring human emotional states by fusing multiple perceptual modalities (such as text, audio, video). Multi-modal emotion recognition can make full use of the complementary between different modalities to improve the accuracy and robustness of emotion recognition. This paper proposes a multi-modal emotion recognition method, however, we briefly outline the single-modal emotion recognition method as some of these techniques can be used for multi-modal recognition.

In text sentiment recognition of natural language processing technology, keyword extraction and machine learning algorithms are used to classify emotions based on text. The paper modelled of Dialogue RNN proposes a new method for recurrent neural networks, which currently process conversations without individually adapting each speaker by tracking the state of each participant throughout the conversation and using this information for sentiment classification. Dialogue GCN uses the inter-dependencies between and within conversation participants to model conversation scenarios and perform sentiment recognition. In speech emotion recognition, we analyze the acoustic characteristics of speech, such as pitch, intensity and sound quality to identify emotions, commonly used audio emotion recognition technologies have sound prosody analysis and speech recognition algorithms. Interaction perception-based attention network (IAAN) proposed the mechanism of attention. Incorporate contextual information into the learned sound representation. In video visual recognition, inferring emotions by analyzing facial expressions, body voice and other visual cues. VAANet proposes to integrate spatial, channel, and temporal attention into a visual 3D CNN and temporal attention into an audio 2D CNN. Based on the combined LEMHI and CNN-LSTM networks. The local network utilizes the LEMHI method to aggregate unidentified video frames into a single frame, which is predicted by CNNs. The global network performs video facial emotion recognition through an improved CNN-LSTM model.

However, single-modal emotion recognition has some disadvantages, including 1)Insufficient information: The input data of a single modality may not provide enough information to accurately identify emotions. For example, non-verbal elements such as tone, intonation and body language which are

critical for emotion recognition may not be captured by words or phrases in text alone. 2) Diversity Challenge: Emotion is a complex concept that involves multiple emotions and emotional states. Single-modal emotion recognition may not capture this diversity because it relies on only one type of data. For example, in text emotion recognition, a single modality may be difficult to distinguish complex emotional states such as birth gas, frustration, and complex emotions. 3) Recognition accuracy limitation: The accuracy of single-modal emotion recognition may be limited by modeling methods and feature extraction methods. There may be differences in the expression of emotions in different modalities and methods of single modalities may not adequately capture these differences, resulting in reduced recognition accuracy.

B. MULTI-MODAL EMOTION RECOGNITION

In recent years, the deep learning of multi-modal emotion recognition has attracted more and more attention. One of the key challenges of multi-modal emotion recognition is how to effectively fuse information from different modalities. The researchers propose various methods to solve this problem, including feature-level fusion, decision-level fusion and model-level fusion. Feature-level fusion combines the features of different modalities, decision-level fusion summarizes the classification results of each modality and model-level fusion models the joint representation of multiple models at the same time.

The feature fusion method is to use feature-level fusion, decision-level fusion and model-level fusion to fuse features from different modes (such as text, video, audio). Ghisal A (2019) first extract features from each modality and combine the features of these modalities layer by layer through the method of layered fusion which improves the performance of sentiment classification through information transfer and fusion between layers. Using the new feature extractor and adaptive margin ternary loss function for training, the model has achieved new achievements in the field of emotion recognition. In order to prevent the problem of large feature dimensions, information synchronization and over fitting between multi-modal information. This paper introduces attention concentration to automatically adjust the attention of the network to effective information. The multi-head attention mechanism is used to fuse the features of audio and video data, which avoids the influence of prior information on the fusion results. The time series of the fusion feature is modeled by the bidirectional gating unit and the auto-correlation coefficient of the time dimension is calculated as the fusion attention. The results show that this attention mechanism can significantly improve the accuracy of emotion recognition. This paper proposes a video multi-modal emotion recognition method based on two-way gating unit and attention fusion which improves the accuracy of emotion recognition in context by applying the dual-gate loop unit and proposes a new network initialization method to further improve the emotion recognition accuracy. In order to solve the weight consistency problem in multi-modal fusion, an attention fusion network is introduced to learn the context information of multi-modality.

C. GRAPH DIALOGUE INFORMATION FUSION

Commonly used models of emotion recognition include feature fusion methods and multi-modal graph neural network methods. Graph neural network method uses graph neural network (GNN) to model dependencies in multi-modal data by propagating information on the graph structure. GNN can capture the correlation between different modalities and provide rich emotional characteristics. GNN and its variants have developed rapidly in text classification, such as Text GCN, Tensor GCN and TextLevelGNN. Recently graph convolutional neural networks have been applied to different multi-modal tasks such as visual dialogue, multi-modal fake news detection and visual question answering. Amir Zadeh (2019) introduced the collection process and annotation procedure of the CMU-MOSEL datasets and studied the architecture and components of the dynamic fusion graph (FGD) model of interpret ability which effectively captures and fuses different modal information for multi-modal language analysis. Jiang applied a novel Knowledge Bridge Diagram Network (KBGN) at a fine-grained level to model the relationship between visual dialogue cross-modal information. Deep learning technology is used to fuse audio and visual cues in the depth model to bridge the emotional gap, fine-tune the per-trained DCNN model through two stages of training and integrate the output in the fusion network to obtain audio-video feature representation. Wang proposes a knowledge-driven multi-modal graph convolution network (KWGCN) for the semantic representation of fake news detection. Khademi introduced the multi-modal Neural Graph Memory Network (MNGMN) for visual question-answering tasks. Ayush Jain (2022) proposed that in conversations involving multiple people, a person's emotions are affected by other speakers and their own emotional state, proposed a context-based graph neural network multi-modal emotion recognition system (COGMEN) and proved the importance of model establishment through detailed ablation experiments. Dependencies and potential correlations between speakers and a unified representation of each speaker through a network of branches. Multi-modal emotion recognition is a method that combines multiple types of input data (such as text, speech, images) to recognize emotions, although it has some advantages but also has the following shortcomings that need to be improved. 1) Extraction of modal interaction relationship of data fusion method: multi-modal emotion recognition needs to fuse data from different modalities which involves feature extraction and multi-modal fusion methods between contexts. Previous studies mainly used the self-attention mechanism or the splicing mechanism between the two modes of data for simple data fusion and the data fusion of the multi-level cross-modal multi-head attention mechanism was not in-depth enough. 2) Difficulty in integrating cross-modal context information: multi-modal emotion recognition needs to comprehensively consider the context information of different modalities. The GAT diagram structure considers both the characteristics of the edge and the characteristics of the integration point. The side refers to the process of the dialogue and the point refers to the content of the dialogue. But how to effectively integrate and leverage this information remains a challenge. However, most of the existing research on emotion recognition focuses on mining the interaction information

between modes and the context information in the dialogue process ignores the correlation information between the role information between multiple modals and the context in the dialogue process.

Aiming at the above problems in multi-modal emotion recognition, this paper uses the multi-head attention mechanism and graph neural network to do information fusion according to the multi-modal fusion problem and the temporal context information extraction problem and proposes a multi-modal fusion dialogue graph attention network (MM DialogueGAT). It effectively solves the problems of modal interaction relationship extraction and cross-modal context information integration of data fusion method in multi-modal emotion recognition problem.

III. MODEL ARCHITECTURE

Self-dependence refers to the fact that the emotional state of a communicator in a conversation may be affected by the previous emotional state of the self. People usually tend to maintain their emotional state. Self-dependence or emotional inertia means that people tend to cling their emotional experiences without being easily disturbed by external circumstances. Interdependence between communicators means that in a conversation, the emotional state of one communicator may be affected by the emotional state of other communicators because people imitate each other's emotions in the process of communication to establish better interaction and resonance. This emotional influence can be transmitted in the conversation and lead to changes in emotional state.

In the process of contextual information modeling of two-person dialogue system or multi-person dialogue system, the emotional dynamics of the communicator in the dialogue are mainly affected by two factors: one is the self-dependence of the communicator and the other is the interdependence between the interlocutor.

Therefore, the cross-modal context information correlation scheme is divided into three ways- sequential coding, speaker-level coding and cross-modal information fusion. The combination of these three different but related context information schemes will create an enhanced context representation to better understand the emotional dynamics in the dialogue system. This combination is conducive to capturing the transmission, interaction and change of emotions in the dialogue, improves the accuracy of emotional recognition.

A. PROBLEM DEFINITION

The framework diagram of the model proposed in the paper is shown in the figure and the whole model is a relatively large architecture which realizes sentiment classification through three modes: text, video, and speech. The model is divided into four different stages, namely sequential context encoding stage, cross-modal context information integration stage, inter-interlocutor graph structure encoder stage, and sentiment classification stage. The model is trained hierarchically, each stage uses the output of the previous stage, and the model is trained using the cross-entropy loss function.

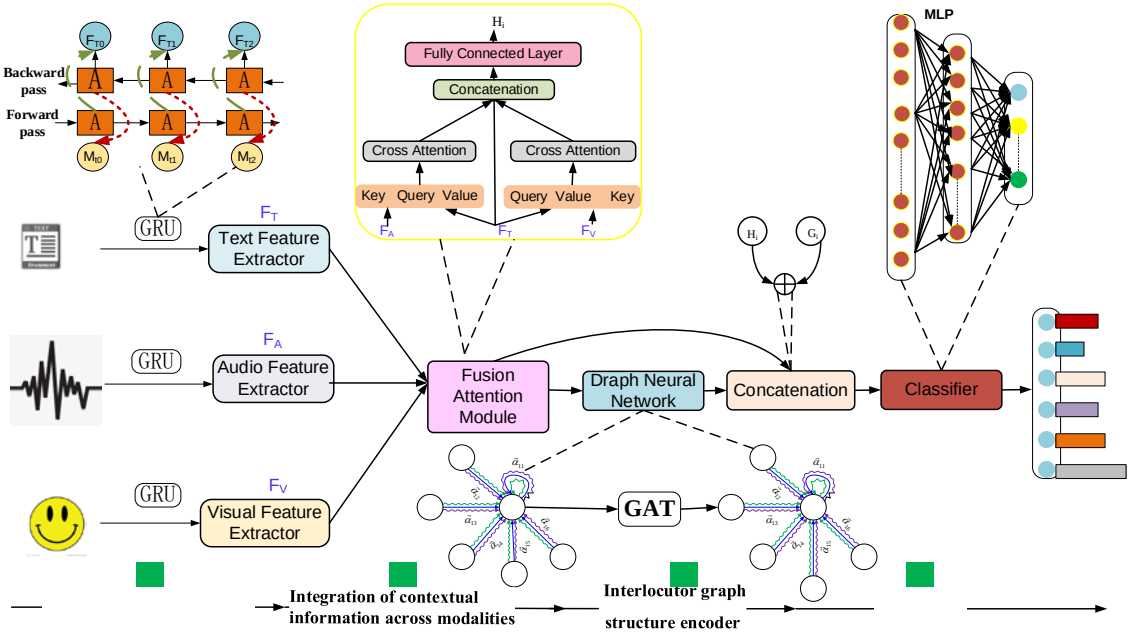
There are 2 interlocutors in a conversation, namely p_i, p_j, one or more utterances composed of words, voices and videos issued by each speaker in the video file, and the task is to predict the emotional label of each paragraph by using the three modal information of text, video, and voice. The emotional categories are happy, sad, neutral, angry, excited, frustrated. Each piece of information consists of text $M_0^t, M_1^t, \ldots, M_n^t$, speech $M_0^a, M_1^a, \ldots, M_n^a$, video $M_0^v, M_1^v, \ldots, M_n^v$. The goal of multi-modal fusion is to solve the problem of temporal context information extraction and transmission by using multi-head attention mechanism and multi-modal dialog graph attention network. The multi-modal information is spoken by the speaker. The multi-modal information M_i spoken by the speaker $p_s(M_i)$.Input for each piece of multi-modal information. The result outputs one of 6 types of emotional data types (Happiness, Sadness, Neutral, Anger, Excited, Frustrated).

B. MM DIALOGUEGAT

In order to fully explore the correlation information between the multi-modal information and the context in the dialogue process and to make the model effectively solve the problem of information jump connection in the context information of the dialogue, this paper uses the multi-head attention mechanism and graph neural network to do information fusion and proposes a multi-modal fusion dialogue graph attention network (MM DialogueGAT). The MM DialogueGAT model contains four key components - sequential context encoding part, cross-modal context information integration part, interlocutor graph structure encoder and sentiment classifier. The overall architecturediagram is shown in Figure 1.

Figure 1 MM DialogeGAT

Figure 1 MM DialogueGAT



1) SEQUENTIAL CONTEXT ENCODING

Conversations are sequential in nature, and contextual information is delivered in sequence. We encode a series of text, audio, and video messages sequentially and input the dialogue into a bidirectional gating unit (GRU) to preserve the context and order of each modality. This step marked number 1 in the Model.

$$F_i^T = \overleftrightarrow{GRU_S}\left(F_{i(+,-)1}^T, M_i^t\right), \text{ for i} = 1, 2, \ldots, \text{N} \quad (1)$$

$$F_i^A = \overleftrightarrow{GRU_S}\left(F_{i(+,-)1}^A, M_i^a\right), \text{ for i} = 1, 2, \ldots, \text{N} \quad (2)$$

$$F_i^V = \overleftrightarrow{GRU_S}\left(F_{i(+,-)}^V, M_i^v\right), \text{ for i} = 1, 2, \ldots, \text{N} \quad (3)$$

M_i^t is a context-free representation of text information; M_i^a is a context-free representation of speech information; M_i^v is a context-free representation of video information; F_i^T is a sequence context-aware representation of text information; F_i^A is the sequence context-aware representation of speech information; F_i^V is the sequence context-aware video information representation, we will express $F_i^T \in \mathbb{R}^{Dt}$ as the feature representation of text information, $F_i^V \in \mathbb{R}^{Dv}$ as the feature representation of video information, and $F_i^A \in \mathbb{R}^{Da}$ as the feature representation of voice information.

2) INTEGRATION OF CONTEXTUAL INFORMATION ACROSS MODALITIES

Most of the existing research on emotion recognition focuses on mining the interaction information between modes and the context information in the dialogue process, but neglects to mine the role information between multi-modal states and the context information in the dialogue process. When processing multi-modal data, the multi-head attention mechanism overcomes the multi-modal information interaction problem by introducing multiple attention heads and learning feature information at different positions and levels, so as to better capture the important features of the input sequence. Each attention head of the multi-head attention mechanism has its own matrix of projected weights for queries, keys, and values, which are learned to capture different feature representations.

Therefore, in the multi-modal information fusion problem, different modal configurations and different combinations use the cross-modal multi-head attention mechanism to establish a multi-head attention layer. Using text, video and audio information as the main and auxiliary modes for information fusion.

The feature vectors of each modal are encoded in a sequential context and captured to become feature vectors F_T, F_V, F_A, these three feature vectors are passed to a multi-head attention module to help the model integrate text, acoustic, and visual information. When calculating attention weights, each header generates a set of weights, which are then weighted and averaged to obtain the final weighted value. Projection is the linear transformation of input Query, Key and Value into different feature spaces. Assuming that each segment of multi-modal information is L in length and dimension is d_k, then each segment of multi-modal information will be transformed into three matrices, $Q_j^T \in R^{N*d_k}, K_j^A \in R^{N*d_k}$, $V_j^T \in R^{N*d_k}$, this article sets F_i^T to Query and Value values to perform multi-head attention operations, visual features F_i^V and audio features F_i^A are used as Key values, respectively, to adjust the attention span

of each segment of the conversation at each time period. Thus, each individual modal is mapped to the text vector space and the corresponding features are connected, transferred to a fully connected layer, and a final feature fusion value is output H_i .The output result H_i used as input values for the next stage.This step marked number 2 in the Model.

$$Query_c^T = F_i^T W_c^Q, Key_c^A = F_i^A W_c^K,$$
$$Value_c^T = F_i^T W_c^V, \ Key_c^V = F_i^V W_c^K,$$
$$c = 1, \ 2, \ 3, \ 4$$
$$P_A = \text{Attn}(Query_c^T, Key_c^A, Value_c^T)$$
$$= softmax\left(\frac{Query_c^T Key_A^T}{\sqrt{d_k}}\right) Query_c^T \quad (4)$$

$$P_V = \text{Attn}(Query_c^T, Key_c^V, Value_c^T)$$
$$= softmax\left(\frac{Query_c^T Key_V^T}{\sqrt{d_k}}\right) Query_c^T \quad (5)$$

$$H_i = \text{Concate}(F_i^T, P_A, P_V) \quad (6)$$

We take text as the main modality, use the emotional representation above the sound to emphasize the important emotional information on the text modality, and then take the text as the main modality and use the visual emotional representation to emphasize the important emotional information on the text modality.

3) INTERLOCUTOR GRAPH STRUCTURE ENCODER

DialogueGAT can be seen as a graph neural network based on the attention mechanism, which is a non-spectral learning method that uses the attention mechanism to calculate the attention weight of each node and its neighbor nodes, and then weighted the node features with the attention weight. This calculation of attention weights usually uses the multi-head attention mechanism to improve the expressiveness and stability of the model. Therefore, GAT can be regarded as a graph neural network based on attention mechanism, which can carry out information propagation and aggregation in graph structure data, and learn and represent node features.

Graph structure(g): In this paper, the output of the multi-headed attention layer is used as the input of the graph structure, and a convolution feature transformation process based on local neighborhood is constructed to create rich speaker-level context feature coding. The graph structure is mainly composed of points, edges, the relationship between points and edges, and the weights of edges, $g = (v, \varepsilon, \gamma, \omega)$.

Point (v): Each segment of multi-modal information in the dialogue represents a point in graph g , and each node v_i sequentially uses the corresponding initialized encoded feature vector H_i, i ∈[1,2,...,N] and we represent the vector H_i as initial features in the graph structure.

Edge(ε): The setting of an edge depends on the dependency of each utterance point in the context information. Each vertex can be connected to all other vertices or connect itself, and if it is performed sequentially, the time complexity required is $O(N^2)$, which is very computationally intensive. To speed things up, we can set the size of past and future windows to reduce the amount of computation. Assuming that the past time window size is wp and the future time window size is wf, the discourse between each utterance vertex n_i and the past time window size wp has edge $v_{i-1}, v_{i-2}, \ldots, v_{i-wp}$, and the discourse between each utterance vertex n_i and the future time window size wf has edge $v_{i+1}, v_{i+2}, \ldots, v_{i+wp}$.

Connections(C_i): Local information constituted of internal dependence and external dependence. For the local information, there are two kinds of relationship between point and edge r_{ij}, one is the external dependency between the speakers, the speaker v_i is set $p_{s(v_i)}$, the speaker v_j is set $p_{s(v_j)}$.The second factor is internal dependence, in the multi-modal information, the order of occurrence with the multi-modal information M_i and $M_j$also affects the construction of the graph. So there are two cases with $i > j$ and $i \leq j$. Global information is all the 8 kinds of connections(C_i). Assuming that there is a dialogue between Q people in a multi-modal information, then in the figure g , there will be a point-edge relationship γ , that is $Q_{v_i} * Q_{v_j} * 2$, when the interlocutor is 2 people, there are the following 8 kinds of point-side relationships. Suppose there are two inter leavers p_i, p_j whose multi-modal information has 6 sentences, which H_1, H_3, H_5 is the p_i speaker's multi-modal information and the H_2, H_4, H_6 is p_j speaker's multi-modal information. The conversation information and diagram are shown below.All of the eight kinds of Connections combined together with globe information.This step marked number 3 in the Model.

(*Note:* $i = j$ *belongs to forword*)

TABLE 1 CONNECTIONS DIAGRAM

Connection	$p_{s(n_i)} \cdot p_{s(n_j)}$	ji: Backword ij: Forword	Diagram[i, j]
C_1	p_i, p_i	Backword	
C_2	p_i, p_i	Forword	
C_3	p_j⊡p_j	Backword	
C_4	p_j⊡p_j	Forword	
C_5	p_i⊡p_j	Backword	
C_6	p_i⊡p_j	Forword	
C_7	p_j⊡p_i	Backword	
C_8	p_j⊡p_i	Forword	

Weight of the edge (ω): In order to more fully represent the characteristics of the node, perform a feature transformation on the node H_i to WH_i , $W \in R^{\tilde{F}*F}$ to map the characteristic dimension F of the node to the dimension $\tilde{F}$. Perform self-attention operation for each node in the graph to calculate the attention weight between any two nodes, and the importance of

node j to node i is calculated as follows:

$$e_{ij} = \alpha[WH_i : WH_j] \quad (7)$$

The H_i is the eigenvector of length F, apply the weighted matrix W to the eigenvectors of the nodes for a linear transformation and calculate the attention coefficient, where i and j are adjacent nodes.

To make the coefficients so far easy to compare at different nodes, they are normalized in the set $\aleph_i$ using the softmax function. In the experiment, the attention mechanism is a single-layer feed-forward neural network with an activation function using LeakyReLU

$$\alpha_{ij} = \frac{\exp(LeakyReLU(e_{ij}))}{\sum_{j\epsilon\aleph(i)} \exp(LeakyReLU(e_{ij}))} \quad (8)$$

Here $\aleph$ is the neighborhood of node i.

$$G_i = \sigma(\sum_{j\epsilon\aleph(i)} \alpha_{ij} W G_j) \quad (9)$$

The learning output characteristics of compute nodes, σ nonlinear transformations.

4) SENTIMENT CLASSIFIER

The feature vector H_i of cross-modal information fusion is G_i stitched together with the feature vector obtained by the speaker-level encoder to obtain the final utterance representation, and then the emotions are classified with a fully connected network and this step marked number 4 in the Model:

$$M_i = [H_i, G_i]$$
$$l_i = ReLU(W_l M_i + b_l) \quad (10)$$

$$p_i = softmax(W_{smax} l_i + b_{2smax}) \quad (11)$$

$$\hat{y}_i = argmax(p_i) \quad (12)$$

IV. MODEL ARCHITECTURE

A. DATASETS

In order to verify the efficiency of the model and to make a fair comparison with the baseline, we do comparative experiments on the IEMOCAP datasets. IEMOCAP datasets are multi-modal data that includes text, visual, and sound data.

IEMOCAP: The IEMOCAP corpus used in this paper is recorded by the University of Southern California, and the datasets contains data from 10 male and female actors during emotional binary interactions. The data includes 9 emotions (angry, happy, excitement, sad, frustration, fear, surprise, other, neural) data and tags. The database contains impromptu sessions and scripted sessions. In total, about 12 hours of audiovisual data were released. For each improvised and scripted recording, the datasets provides detailed audiovisual and textual information including audio and video for both interlocutors. In addition, for each statement of the recording, annotations are provided for classification and dimension labels from multiple annotators. The datasets uses motion capture and audio/video recording data, which is collected from 5 binary sessions across 10 topics. In cross-validation, we use the first four dialogs as the training set and the last dialog as the test set. Each session consists of a different dialogue in which 1 male and 1 female actor executes the script and participates in spontaneous, impromptu conversations triggered by emotional scene cues. The corpus has a total of 10,039 audios with a total duration of nearly 12h, and each audio also contains video, face change capture, and text information. This article covers 6 types of emotional data for 9 emotions, including (Happiness, Sadness, Neutral, Anger, Excited, Frustrated).

TABLE 2 THE WAY IMEOCAP DEVIDED

datasets	Split	Utterance	Dialogue
IMEOCAP	Train/Val	5810	120
	Test	1623	31

TABLE 3 THE DIMESION SIZE OF THE DATASETS

	IMEOCAP
Acoustic	100
Visual	100
Textual	512

B. BASELINES

DialogueRNN: A relatively advanced method in the field of emotion recognition. It is a recursive network that uses two gated recurrent units (GRUs) to track the state of each speaker in a conversation.

DialogueGCN: This is a graph-based neural network-based way of sentiment recognition that focuses on text.

MMGCN: This is a multi-modal AER approach that uses GCN models to fuse multi-modal information interactions.

COGMEN: Modeling complex dependencies based on GNN architecture using local information (i.e., mutual internal dependencies between speakers) and global information (i.e., context information).

DIMMN: This is a multi-modal interaction model that mines cross-modal dynamic dependencies between different groups in the dynamic process.

In addition, we also investigate the different configurations for the proposed models.

MM DialogueGAT(Without Fusion):This is a model which have the step of 1.Sequential context encoding,3. Interlocutor graph structure encoder and 4.Sentiment classifier.

MM fusion: This is a model which have the step of 1. Sequential context encoding,2.Integration of contextual information across modalities and 4.Sentiment classifier.

MM DialogueGCN(Without Fusion):This is a model which have the step of 1.Sequential context encoding,3. Interlocutor graph structure encoder and 4.Sentiment classifier.In the step 3,the model GCN instead of GAT.

MM DialogueGCN: This is a model which have the step of 1.Sequential context encoding,2.Integration of contextual information across modalities. 3.Interlocutor graph structure encoder and 4.Sentiment classifier.

In the step 3, the model GCN instead of GAT.

MM DialogueGAT: This is a model which have the step of 1.Sequential context encoding,2.Integration of contextual information across modalities. 3.Interlocutor graph structure encoder and 4.Sentiment classifier.

V. EXPERIMENTAL RESULTS

Tabular comparison of IEMOCAP datasets results with baseline methods; Acc = accuracy; Bold results are the best performance;

TABLE 4 COMPARISON WITH THE BASELINE METHODS ON IEMOCAP

Models	IEMOCAP : Emotion Categories							
	Happiness	Sadness	Neutral	Anger	Excited	Frustrated	Avg(Acc)	Avg(F1)
Dialogue RNN	32.80	78.00	59.10	63.30	73.60	59.40	63.30	62.80
Dialogue GCN	42.75	84.54	63.54	64.2	63.08	66.99	65.25	64.18
MMGCN	33.18	66.96	56.03	63.9	68.14	58.51	60.1	59.29
COGMEN	51.90	81.70	68.60	66.00	75.30	58.20	68.20	67.60
DIMMN	30.20	74.20	59.00	62.70	72.50	66.60	64.70	64.10
MM Dialogue GCN	86.11	89.61	77.05	74.3	91.54	74.41	79.18	79.35
MM Dialogue GAT	89.05	91.87	80.17	76.5	88.99	77.25	**81.46**	**81.99**

DATESET

Avg(Acc)=Accuracy average ; Avg(F1)=F1 average

A. COMPARING STUDY

To verify the effectiveness of the proposed network , the experimental results show that the proposed in this paper has excellent performance in the recognition accuracy of six different emotions, including recognition accuracy rate (Acc.) and the weighted average F1.

The MM DialogueGAT model improves the Acc. and F1 by 18.16% and 19.19% compared with DialogueRNN. We anaylze the reason is that MM DialogueGAT not only considers the sequential propagation of context information, but also considers the connection problem of hop information. Each node can propagate and receive information from multiple hops, each node can capture information from nodes at different distances, enabling the implementation of skip connections for information. DialogueRNNs use gated recurrent unit (GRU) networks to model individual speaker states. In contrast, the DialogueRNN model can only process contextual information between interlocutors for conversations with a large number of utterances.

MM DialogueGAT model improves by 16.21% and 17.81% in Acc. and F1 respectively.We anaylze the reason is that MM DialogueGAT performs multi-modal information representation of text, speech, and video in sequence context perception, the multiple data sources providing a more comprehensive and rich set of information. DialogueGCN is a graph neural network-based emotion recognition that only focuses on text. Single-modal information can lead to poor performance in fine-grained emotion recognition.

The MM DialogueGAT model improves by 21.36% and 22.7% in Acc. and F1. The reason is that MM DialogueGAT dynamically assigns the weights of different neighbor nodes to each node. This allows the model to focus at each step on neighbor nodes that are more meaningful to the task at hand. Dynamic adjustment of edge weights allows the model to finely control between different edges, striking a balance between capturing local information and global information. MMGCN is a multi-modal AER method that uses the GCN model to fuse multi-modal information interactions without dynamically assigning weights to different neighbor nodes.

The MM DialogueGAT model improved the Acc. and F1 by 13.26% and 14.39%. The reason is that MM DialogueGAT not only considers the internal dependencies between speakers and the context global information, but also considers the internal and external dependencies between the speakers to form complex dependencies. COGMEN only uses local information (mutual internal dependencies between speakers) and global information (context information) to model complex dependencies based on GNN architecture. The modeling process neglected to consider the internal and external dependencies between speakers to form complex dependencies.The MM DialogueGAT model improved the Acc. and F1 of the model 16.76% and 17.89%.The reason is that DIMMN is a multi-modal interaction model that mines the cross-modal dynamic dependence between different groups in the process of mining dynamics and only considers the information interaction between modes.

B. ABLATION STUDY

TABLE 5. ABLATION STUDIES BASED COMPARISON TO VALIDATE THE IMPACT OF EACH PART OF THE MODEL

Model	IEMOCAP	
	Acc.	F1
MM DialogueGAT(Without Fusion)	76.33	76.49
MM Fusion	76.88	77.12
MM DialogueGCN(Without Fusion)	75.34	75.51
MM DialogueGCN	79.18	79.35
MM DialogueGAT	81.46	81.99

In order to better understand the contributions of different modules in the multi-modal Fusion Dialogue Graph Attention Network (MM DialogueGAT), we conducted multiple ablation experiments on the IEMOCAP datasets. The corresponding results were compared with the Acc. and F1 on the IEMOCAP datasets.

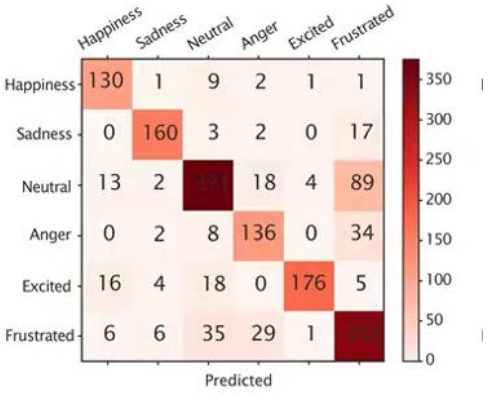


Figure 2. MM Fusion

In MM Fusion, after the graph structure encoder step between the interviewers is deleted, the model can only connect the information of the 6-cephalic attention extends the attention mechanism into multiple parallel attention heads, each of which can learn different relationships. This can better capture the

diverse relationships between nodes and improve the expressiveness of the model. As a result, the Acc. and F1 of the model decreased by 4.58% and 4.87%. The result shows in figure 2.

In the MM DialogueGCN(Without Fusion), after the cross-modal context information integration step in the model is deleted, the model does not have the process of jumping and connecting multi-modal information.In the emotion recognition task, different levels of feature representations may contain different levels of semantic information. Through information jump connections, the underlying and deep features can be fused together to better capture the multi-level emotional information in the multi-modal information. The accuracy of emotion recognition may depend on subtle differences in multi-modal information, such as the position of emotional words and grammatical structure. Information hopping connection helps to retain the detailed information in the underlying features which is very important for more accurate emotion recognition, so after the cross-modal context information integration step is cut, the Acc. and F1 of the model decrease by 6.12% and 6.48%. The result shows in figure 3.

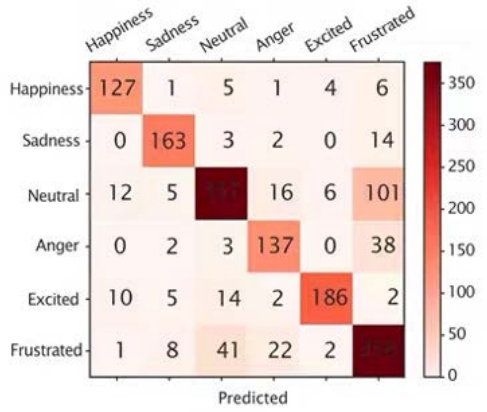


Figure 3. **MM DialogueGCN(Without Fusion)**

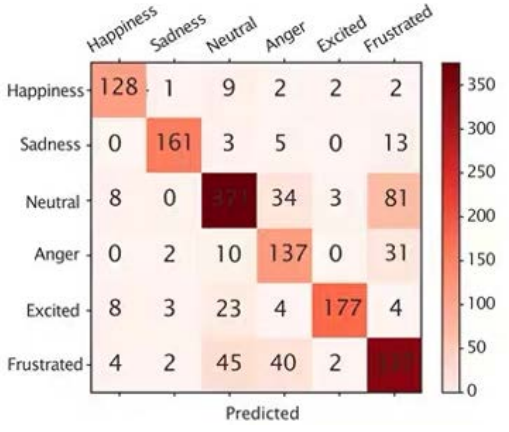


Figure 4. **MM DialogueGAT(Without Fusion)**

In the MM DialogueGAT(Without Fusion), The graph attention network can learn the embedding or representation of nodes, which can capture the context of nodes, neighboring nodes, and the structure of the entire graph, which helps to extract the semantic information of nodes, so as to effectively capture the complex relationships and connections between nodes. Moreover, graph attention networks use connections between nodes to pass and aggregate information. This approach enables the graph attention network to pass local information to the global scope, resulting in a better understanding of the context of the nodes. The Acc. and F1 of the model decrease by 5.13% and 5.5%. The result shows in figure 4.

In the MM DialogueGCN model, after changing the graph structure of the third part between the interlocutors in the model to GCN, the accuracy rate (Acc) and weighted F1 decreased by 2.28% and 2.64% respectively.

The analysis of the experimental results reveals that the recognition accuracy for the 'Excited', 'Happiness', and 'Sadness' categories is relatively high, while the recognition accuracy for the 'Neutral' category is low. This directly results in sub-optimal performance in multi-modal emotion recognition. The primary reason for this is the lack of distinct multi-modal features for the 'Neutral' category. The result shows in figure 5 and 6.

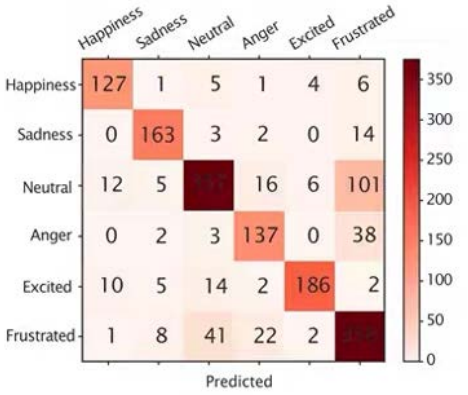


Figure 5. **MM DialogueGCN**

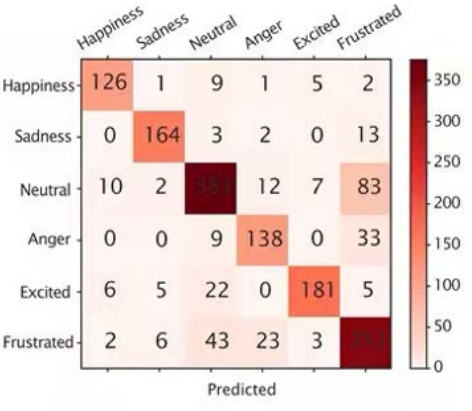


Figure 6. **MM DialogueGAT**

The time window size in sentiment recognition refers to how long a time range is selected in the text data for sentiment analysis. The choice of time window size is important in emotion recognition tasks because it can affect the sensitivity of the model to emotional changes and its ability to understand context. A longer time window can provide more contextual information and help the model better understand the context and context of the text. And emotions usually change over

time. A longer time window can help the model capture trends in the evolution of emotions. But if sentiment analysis requires timely response, a shorter time window may be more suitable, as it captures near-term emotional changes more quickly. Longer time windows can increase computational complexity, especially for large-scale, multi-modal information data. Therefore, the appropriate time window size will greatly improve the recognition efficiency of the model. From the above table, we can see the influence of different past context window sizes (WP) and future context window sizes (WF) on the output results of Model DialogueGAT. Experimental results show that when WP=WF=10, the model recognition rate reaches the highest. With the increase of the size of the past context window and the future context window, the accuracy rate of the model does not increase proportionally, but when the past context window and the future context window are greater than 10, the effective information fused in the multi-modal information will also decrease and the model recognition efficiency will decline.The result shows in figure 7.

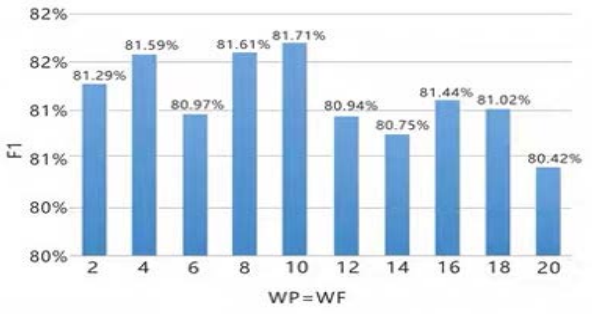


Figure 7. Effect of time window size on f1

VI. DISCUSSION

The results validate that the approach proposed in this paper improves the average accuracy by up around 15%. Model efficiency is improved for several reasons. One is that the model benefits from rich information sources and utilizes model fusion techniques to integrate information across different modalities. The second reason is that the model not only effectively captures contextual information but also acquires information with skip connections. The model can efficiently process and extract features from multi-modal data, enhancing emotion recognition efficiency.

The results of the ablation experiments reveal that the process of fusing multi-modal information plays the most crucial role. Secondly, the graph neural networks also plays an important role. Using only a fusion structure without a graph structure have following problems: 1)Information loss. The model struggle to capture complex dependencies between different modalities. This could result in the loss of crucial information and hinder the model's ability to understand the relationships between various data sources. 2).Insufficient context: A fusion structure alone might not provide sufficient contextual information, potentially leading to an incomplete understanding of emotional states. Contextual cues are often vital for accurate emotion recognition. 3).Skipping connections: emotional states may be related to non-linear dependencies that are more easily captured using a graph structure. A fusion structure might not effectively handle such non-linear relationships.On the other hand, when using only a graph structure without a fusion structure, there have following issues:1).Information isolation problem: Graph structures primarily focus on capturing relationships and dependencies between data points, potentially overlooking important information between different modalities or features.2).Dimensional challenge: Representing relationships between data points using a graph structure can lead to high-dimensional data, which can increase computational complexity and reduce model efficiency.3).Feature extraction difficulty.Therefore, it is necessary to strike a balance both fusion and graph structures to better handle multi-modal data and capture emotional states effectively.

In MM DialogueGAT model, there is a need for further enhancement of the recognition accuracy for the ‘Neutral’ emotion. ‘Neutral’ emotion typically lacks distinct emotional features because it lacks strong emotional states. This makes feature extraction for "Neutral" emotion more challenging in audio, text, or video data. Additionally, the certainty of emotion labels can be a challenge in itself. Different individuals may assign different emotion labels to the same emotional events, making the recognition of ‘Neutral’ emotion complex. Some emotional events might be considered to have mild emotions and not clearly defined as ‘Neutral’. Therefore, how to label and recognize such emotions more accurately and clearly is an important research direction.

In MM DialogueGAT model, both overly large and overly small time windows can impact the efficiency of emotion recognition. When the time window is too large, instantaneous emotional changes might be smoothed or neglected, resulting in the loss of crucial emotional information. This can introduce temporal offsets and hinder adaptation to rapid emotional changes. Conversely, smaller time windows can lead to challenges in labeling data when emotions change frequently, making it susceptible to noise interference, thereby affecting recognition efficiency.

VII. CONCLUSION

In this paper, we uses the multi-head attention mechanism and graph neural network to do information fusion and proposes a multi-modal fusion dialogue graph attention network (MM DialogueGAT). This paper aims to improve the understanding of the context in conversation, so as to improve the effect of emotion monitoring under multi-modal information. The model uses a multi-head attention layer solves the problem of information jump connection in the temporal context, Finally, the local information composed of internal and external dependencies between the speakers and the global information composed of context information uses the combination of the multi-head attention mechanism and the graph attention network to model the complex dependencies in the dialogue.The future research direction will set the parameters of the graph neural network model (MM DialogueGAT) in this paper more reasonably and extend the model to the dialogue scenario of multi-person communication so that the model is more universal in the process of emotion detection. In addition, the performance of the model in recognizing the neutral category is sub-optimal which is attributed to the lack of prominent features for the neutral category in the current emotion recognition models. In future research, we will further refine the MM DialogueGAT

model by incorporating feature extraction methods for the multi-modal network structure. Finally, we plan to use the classification results obtained by the model to generate emotional correspondence, so that human-computer dialogue can achieve more fine-grained emotional generation.

ACKNOWLEDGMENT

Rui Fu and Xiaomei Gaicontributed to the main idea and the methodology of the research. Rui Fu and Mohammed Abdulhakim Al-Absi designed the experiment, performed the simulations, and wrote the original manuscript. Meng Jiang and Ye Li contributed significantly to improving the technical and grammatical contents of the manuscript. Xuewei Wang and Mohammed Alam reviewed the manuscript and provided valuable suggestions to further refine it. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] P. Kumar, B, Rama, ''A BERT based dual-channel explainable text emotion recognition system.'' *Neural Networks*, vol.150, pp.392-407, 2022, doi: 10.1016/j.neunet.2022.03.017.

[2] N. Majumder, D. Navonil, S. Poria, P. Soujanya, ''Dialogue RNN: An Attentive RNN for Emotion Detection in Conversation.'' *Archives*, vol.37, pp.2-14, 2019, doi: 10.48550/arXiv.1811.00405.

[3] G.Deepanway, M.Navonil, P.Soujanya, ''Dialogue GCN:A Graph Convolution Neural Network for Emotion Recognition in Conversation,'' *ACL Anthology*, vol.37, pp.154-164, 2019, doi: 10.18653/v1/D19-1015.

[4] Z.Suping, J.Jia, Q.Wang, D.Yufei, ''Inferring Emotion from

[5] Conversational Voice Data:A Semi-Supervised Multi-Path Generative Neural Network Approach.'' *Proceeding of the AAAI Conference on Artificial Intelligence*, vol.32, pp.579-586, 2019, doi: 10.1609/aaai.v32i1.11280.

[6] S.Linyeh, L.Yunshao, L.Chichun, D.Yufei,''An Interacion-aware Attention Network for Speech Emotion Recognition in Spoken Dialogs.'' *ICASSP*, vol.32, pp.125-135, 2019, doi: 10.1109/ICASSP.2019.86683293.

[7] Z.Sicheng. M.Yunsheng, G.Yang, Y.Jufeng. ''An End-to-End Visual-Audio Attention Network for Emotion Recognition in User-Generated Videos.'' *Archives*, vol.37, pp.18-28, 2020, doi: 10.48550/arXiv.2003.00832.

[8] H.Min, W.Haowen, W.Xiaohua, Y.Juan, ''Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM network.'' *Journal of Visual Communication and Image Representation*, vol.59, pp.176-185, 2019, doi: 10.1016/j.jvcir.2018.12.039.

[9] V.Monu, V.Santosh Kuamr, S.Girdhari, M. Subrahmanyam.''LEARNet: Dynamic Imaging Network for Micro Expression Recognition.'' *IEEE Transactions on Image Processing*, vol.59, pp.1618-1627, 2019, doi: 10.1109/TIP.2019.2912358.

[10] C.Vishal, K.Purbayan, G.Ashish, S.Nirmesh.''M2FNet:Multi-model Fusion Network for Emotion Recognition in Conversation'' *Computer Science*, vol.15, pp.158-167, 2022, doi: 10.48550/arXiv.2206.02187.

[11] T.Guichen, X.Yue, L.Ke,L.Ruiyu.''Multi-modal emotion recognition from facial expression and speech based on feature fusion.'' *Multimedia Tools and Applications*, vol.82, pp.16359-16373, 2022, doi: 10.1007/s11042-002-14185-0.

[12] H.Ruohong, S.Jia, B.Shenglin, L.Ronghua.''Video multi-modal emotion recognition based on BiGRU and attention fusion.'' *Multimedia Tools and Applications*, vol.80, pp.8213-8240, 2021, doi: 10.1007/s11042-002-14185-0.

[13] Z.Shiqing, Z.Shiliang, H.Tiejun.''Multi-modal Deep Constitutional Neural Network for Audio-Visual Emotion Recognition.'' *ICMR*, vol.80, pp.281-289, 2019, doi: 10.1145/2911996.2912051.

[14] R.Minjie, H.Xiangdong, L.Wenhui, S.Dan.''LR-GCN:Latent Relation-Aware Graph Convolutional Network for Conversational Emotion Recognition.'' *IEEE Transactions on Multimedia*, vol.24, pp.4422-4432, 2021, doi: 10.1109/TMM.2021.3117062.

[15] E.Krumbuber, S.Lina, L.Karen.''The role of facial movements in emotion recognition.'' *Nature reviews psychology*, vol.24, pp.283-296, 2023, doi: 10.1038/s44159-023-00172-1.

[16] L.Shuai, G.Peng, L.Yating, F.Weina.''Multi-modal fusion network with complementary and importance for emotion recognition.'' *Information Sciences*, vol.619, pp.679-694, 2023, doi: 10.1016/j.ins.2022.11.076.

[17] Baevski. A, Y. Zhou, Mohamed, A, Auli, M. ''wav2vec 2.0: a framework for self-supervised learning of speech representations.'' *Neural Inf Process*.vol.33,pp.12449-12460.2020,doi: 10.5555/3495724.34 96768.

[18] H.Chang, Y.Zong, W.Zheng, C.Tang, J.Zhu, and X.Li. '' Depression assessment method: an EEG emotion recognition framework based on spatiotemporal neural network. Front.''*Psychiatry*.12,vol33, pp.2620-2630. doi: 10.3389/fpsyt.2021.837149.

[19] H.Chang, Y.Zong, W.Zheng, Y.Xiao, X.Wang, J.Zhu. ''EEG-based major depressive disorder recognition by selecting discrimination features via stochastic search.'' *J. Neural Eng*.vol.20, pp.2620-2630, 2023, doi: 10.1088/1741-2552/acbe20.

[20] W.Binqiang, D.Gang, Z.Yaqian, L.Rengang, C.Qichun.''Hierarchically stacked graph convolution for emotion recognition in conversation.'' *Knowledge-Based Systems*.vol.263, pp.215-225, 2023, doi: 10.1016/j.knosys.2023.110285.

[21] L.Dayu, Z.Xiaodan, L.Yang, W.Suge.''Enhancing emotion inference in conversations with commonsense knowledge.'' *Knowledge-Based Systems*.vol.263, pp.1074-1085, 2021, doi: 10.1016/j.knosys.2021.107449.

[22] G.Qinquan, Z.Hanxin, L.Gen, T.Tong.''Graph Reasoning-Based Emotion Recognition Network.'' *IEEE Access*.vol.9, pp.6488-6497, 2021, doi: 10.1109/ACCESS.2020.3048693.

[23] F.Hayforrd, X.Xiaofen, G.Kailing,X.Xiangmin.''Emotion Recognition With Knowledge Graph Based on Electrode Activity.''*Frontiers*.vol.16, pp.178-189, 2022, doi: 10.3389/fnins.2022.911767.

[24] L.Huating, X.Xiaodong, W.Miao.''Sparse Spatial-Temporal Emotion Graph Constitutional Network for video Emotion Recognition.''*Computational Intelligence and Neuroscience*.vol.1, pp.1123-1134, 2022, doi: 10.1155/2022/3518879.

[25] K.Maryam, S.Akane.''Exploiting social graph networks for emotion prediction.'' *Scientific reports*.vol.13, pp.6069-6080, 2023, doi: 10.1038/s41598-023-32825-9.

[26] Z.Duzhen, C.Feilong,C.Xiuyi.''DualGATs: Dual Graph Attention Networks for Emotion Recognition in Conversations.'' *Association for Computational Linguistics*.vol.1, pp.7395-7408, 2023, doi: 10.18653/v1/2023.acl-long.408.

[27] W.Binqiang, D.Gang, Z.Yaqian, L.Rengang.''Hierarchically stacked graph convolution for emotion recognition in conversation.'' *Knowledge-Based Systems*.vol.263, pp.735-748, 2023, doi: 10.1016/j.knosys.2023.110285.

Rui FU is a Professor in the institute of intelligen t manufacturing at Weifang University of Scienc e and Technology, China. She received her (MS) degree in System Theory from Qingdao Universit y -China in 2012-2015. She earned her Ph.D. degree in the Department of Information and Communication Engineering at Dongseo University, Kor ea. Her research interests include Artificial Intelligence, VANET, UAVs/ Drone, Deep learning, Image Detection and Mathematics.

Xiaomei Gai was born in Weifang , Shandong Province, China, in 1987. She received the Master degree from Dalian Maritime University. She earned her Ph.D. degree in Wonkwang Univeraity, Korea. Now, she works in Weifang University of Science and Technology, Her research interests include Computer Language teaching.

Mohammed Abdulhakim Al-Absi received his B S degree in Computer Application from Bangalore University in India. He earned his (MS) degree at Dongseo University- South Korea in 2018. He finished his PhD from the Department of Information and Communication Engineering at Dongseo University2023, Currently, he is a researcher professor at Dongseo University. His research interests include IoT, VANET, UAV, AI, Crypt-ology, Network Security, Side-Channel Attack, , Deep learning, Cloud computing, Computer Networks and Digital Communications.

Ahmed Abdulhakim Al-Absi is an associate Professor in Smart Computing department, Kyungdong University Global Campus, South Korea. Dr. Al-Absi received his Ph.D. degree in Computer Science (specializing in Big Data Processing) from Dongseo University, Korea. He received his Master of Computer Science degree from University Utara Malaysia- Malaysia in 2011, and a bachelor's degree in computer applications from Bangalore University- India in 2007. He has more than ten years of experience in teaching and university lecturing in the areas of database design and computer algorithms. In the field of research and publication, Dr. Al-Absi's has published numbers of research papers in peer-reviewed international journals and conferences. His research interests include Database Systems, Big Data, Hadoop, Cloud computing, Distributed systems, Parallel computing, High-performance computing, VANET, and bioinformatics.

Mohammad Alam holds a Ph.D. degree in Computer Science from the University of Aveiro, Portugal and has an MS degree in Computer Science from International Islamic University Islamabad, Pakistan. In 2009, he joined the Instituto de Telecomunicações - Aveiro (Portugal) as a researcher and completed his Ph.D. from the University of Aveiro with a specialization in InterLayer and Cooperative Design Strategies for Green Mobile Networks.He is working as Senior lecturer at London South bank University. His research interests include IoT, Realtime wireless communication, 5G, Vehicular network, Context-aware systems, and Radio resource management in next-generation wireless networks.

Ye Li received the M.S. degree from Ocean University of China, Qingdao, China. At present, he is a lecturer of Weifang University of science and technology in China. His current research interest is machine learning and computer vision,Computer Networks and Digital Communications.

Meng Jiang is a Professor and Dean of the College of Language Intelligence at Sichuan International Studies University, China. He is a Ph.D. student supervisor and reviewer of many academic journals. He received his Ph.D. degree in the School of Foreign Languages at Shanghai Jiaotong University, China. His research interests include second language conceptual acquisition, cognitive neurolinguistics, language pathology, and language intelligence.

Xuewei Wang received his BS from Naval Aeronautical University, Yantai, PR China, in 1995; MS from Naval Aeronautical University, Yantai, PR China, in 1998; Ph.D. from Beihang University, Beijing, PR China, in 2005. He is now a professor at the Computer College, Weifang University of Science and Technology, Weifang, PR China. His recent research interests include digital image processing and computer vision.