Neurocomputing 569 (2024) 127109

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition

Tao Meng [a], Yuntao Shou [a], Wei Ai [a,*], Jiayi Du [a], Haiyan Liu [b], Keqin Li [c]

[a] College of Computer and Information Engineering, Central South University of Forestry and Technology, Hunan, China
[b] College of Information Engineering, Changsha Medical University, Hunan, China
[c] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

ARTICLE INFO

Communicated by Z. Tu

Keywords:
Emotion recognition in conversation
Heterogeneous Graph Neural Network
Multi-messaging
Self attention mechanism

ABSTRACT

As an important development direction of natural language processing, emotion recognition in conversation (ERC) remains a challenge in sentiment analysis. Given the large-scale dialogue datasets and their wide application in the fields of recommendation systems and human–machine dialogue systems, researchers have begun to pay more attention to the issue of ERC. In recent research, the task of ERC has been largely based on the graph structure to model the speaker level. However, most existing studies simply splice multimodal features, and the heterogeneity of multimodal features tends to be overlooked. Hence, this paper proposes a multivariate messaging framework to embed heterogeneous information into multimodal relational graphs. In the process of aggregating graph node information, we take into account the homogeneity of nodes and assign different weights to different nodes so as to better aggregate semantic information. In order to improve the robustness of the model, we utilize the mechanism of sharing weights among neighbors to reduce the number of network parameters and improve the generalization ability of the model. In so doing, the node information is aggregated through the constructed graph network, and the final semantic vector representation is obtained. Experiments over two benchmark datasets for ERC show that our proposed model achieves improved performance in accuracy and F1 value.

1. Introduction

Although significant advances have been made in deep learning research, ERC, an important branch of natural language processing, remains a challenging task. Over the past decades, social networking sites such as Twitter, Meta, YouTube, and Reddit have revolutionized the way people communicate and become large social conversation datasets. This has drawn more and more researchers to focus on the ERC and relevant studies. Taking Fig. 1 as an example, each speaker in ERC contains three modal features: video, audio, and text. Our task is to use these three modal information to accurately identify the speaker's emotions. ERC can help machines understand the emotional changes of human beings in the process of communication and produce corresponding empathetic responses. Therefore, ERC has been widely used in spam blocking, health care, advertising, recommendation systems, opinion mining, human–machine dialogue systems, and other fields. However, the ability on the part of the models to accurately understand the emotions behind an expression and respond accordingly remains a significant challenge in sentiment analysis. For example, there is heterogeneity between different modal features, i.e., significant differences between different modal features. We must eliminate the differences between modalities in multi-modal feature fusion to exploit the complementary semantic information between modalities fully.

In a recent study of ERC tasks, Ghosal et al. [1] employed graph convolutional networks (GCN) to examine self and inter-speaker dependency. Taichi et al. [2] used the relation-aware graph attention network (RGAT) to assign different weight vectors to different relationship nodes. They also introduced relational position-encoding vectors to provide contextual information about the graph structure. Zheng et al. [3] simulated human interaction in a conversational context through GGCN and introduced a multi-head attention mechanism to calculate the weightings of the context vectors. Although the application of graph neural networks (GNNs) can effectively model the speaker-level context, the aforementioned methods ignore both the heterogeneous network nodes in relation graphs and the differences between different relations. However, eigenvectors between different modalities are heterogeneous, and so are graph structures formed by

* Corresponding author.
E-mail addresses: mengtao@hnu.edu.cn (T. Meng), yuntaoshou@csuft.edu.cn (Y. Shou), aiwei@hnu.edu.cn (W. Ai), dujiayi@csuft.edu.cn (J. Du), liuhy_csmu@163.com (H. Liu), lik@newpaltz.edu (K. Li).

https://doi.org/10.1016/j.neucom.2023.127109
Received 10 January 2023; Received in revised form 6 September 2023; Accepted 6 December 2023
Available online 12 December 2023

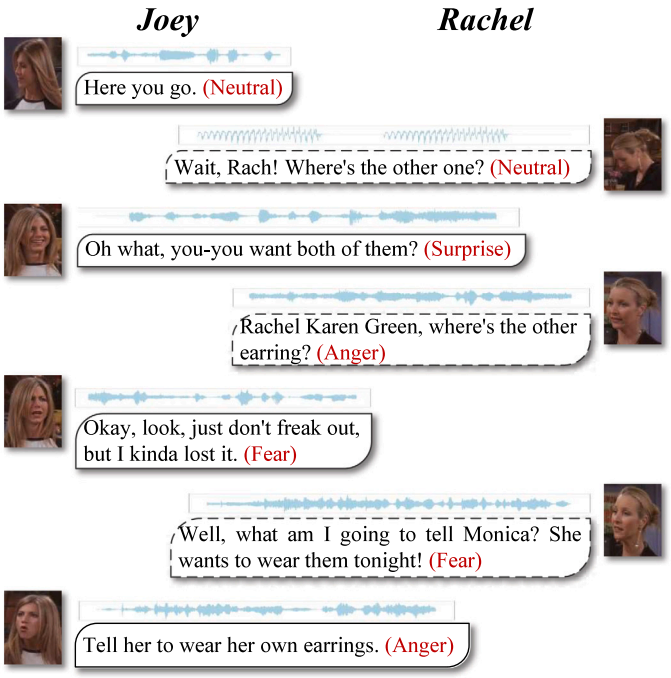


Fig. 1. An example of conversational emotion recognition from the IEMOCAP dataset. Each sentence contains three modal information, i.e., text, video, and audio. The task of multimodal emotion recognition is to use multimodal information to identify the speaker's emotional state.

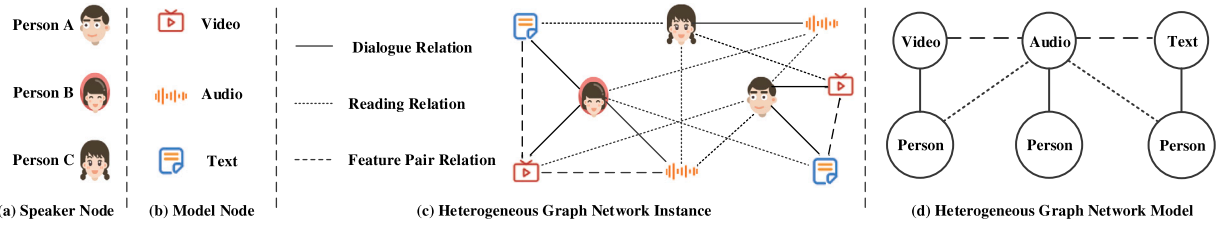


Fig. 2. Illustration of a multimodal heterogeneous graph. (a) Speaker type nodes (i.e., Person A, Person B, Person C). (b) Three different types of modal nodes (i.e., Video, Audio, Text). (c) A multimodal heterogeneous graph network includes four types of nodes and three types of relations. (d) A description of the relational pattern of the heterogeneous graph network.

eigenvectors of different modalities [4]. Heterogeneous graphs, which are used to represent composite relationships between objects, contain rich semantic information. As shown in Fig. 2(c), the heterogeneous graph illustrates the relationship between different modes and nodes. Specifically, in a specific dialogue, the speaker expresses his emotions by sending various modal information, and there is a dialogue relationship between the speaker and the modal information. In addition, other speakers may read the dialogue information and have an important impact on their own emotions. Therefore, other speakers may have reading relationships with modality information. Finally, the speaker may send information (such as video) with modal features such as text, voice, and image to strengthen his emotional expression, and there is a feature pair relationship between multi-modal information. It can be seen that there is rich semantic information between different types of nodes in the dialogue heterogeneous graph. Currently, increasing research [5,6] evidence shows that considering heterogeneity not only improves the ability to capture semantic information, but also improves the accuracy of node representation. In view of this, this paper argues that taking into account the heterogeneity between different modes will enhance the model's ability to understand the conversation from different perspectives.

Heterogeneous graphs generally refer to graphs with different types of nodes or nodes of the same type with different properties, which have been widely used in a variety of natural language processing tasks. For instance, Yao et al. [5] used graph convolutional neural networks to aggregate information from different domains. Sheng et al. [6] modeled the contextual and phrase-level semantic features of the discourse through heterogeneous GNNs. Nonetheless, utilizing heterogeneous GNNs to obtain rich contextual and semantic information and how to further deal with heterogeneous information is still a relatively new field of ERC. At the same time, existing GNNs tend to perform better in shallower neural networks, and as networks become deeper, there will be serious over-smoothing. Also, GNNs show some invalidity when dealing with heterogeneous nodes.

For the task of ERC, as most current models in existing studies do not take into account the heterogeneous characteristics of network nodes, this paper designs a multivariate message transmission model for the emotion recognition task, which comprehensively examines the different types and relationships of network nodes. Our model assigns a weight to cope with the effects of different relations. In addition, this study also proposes an algorithmic mechanism of sharing weights among neighbor nodes in order to reduce model complexity and make the model more robust. Specifically, we first extract and encode sentence words, speech signals, and image regions simultaneously [7]. Secondly, the obtained text, speech and image coding features are input into Bi-LSTM to model the context history information, and then use the dialogue, reading and feature pair relationships to construct a multimodal heterogeneous graph. In this way, the speaker relationship and semantic information in the dialogue can be well preserved. Furthermore, we design a Multivariate Message Passing Graph Convolutional

Network (MMPGCN) to better fuse the multivariate information between relations and capture the correlation between different relations. In MMPGCN, we define the concept of network node homogeneity rate, which is used to measure the degree of homogeneity of each node in the network, so as to determine the feature difference between different nodes. Given the large differences between the nodes [8], this study measures the influence of each edge in the nodes of a graph based on the homogeneity rate between different modalities, and then the learnable weight is multiplied by the homogeneity rate as the final edge weight. Finally, the semantic information obtained by the MMPGCN model is passed through the fully connected layer and the activation function to obtain the classification result of the emotional label.

The main contributions of our work are as follows:

1. This paper proposes a new algorithmic model, called MMPGCN, which comprehensively analyzes the semantic information of the dialogue context, the relationship between speaker levels, and the heterogeneity between multimodal nodes. Furthermore, the model can be used to detect emotions in multiple rounds of conversations.
2. This study defines the concept of node homogeneity based on the different relationships between nodes, which is used to measure the degree of homogeneity of each node within a network structure to determine the difference in features between different nodes. Based on the concept of node homogeneity, the weights of each node are dynamically determined. To improve the generalization ability of the model, based on the idea of neighbors sharing weights, a mechanism to reduce the number of parameters in the network is proposed.
3. To the best of our knowledge, this is the first study to apply multimodal heterogeneous information to ERC, which builds a GNN to simulate the interaction between speakers based on the heterogeneity of different nodes.
4. Experiments on two publicly available benchmark datasets show that the model proposed in this study has improved accuracy in recognizing multiple emotion labels.

The remainder of this paper is organized as follows: Section 2 summarizes relevant studies on emotion recognition; Section 3 describes ERC in mathematical language and elucidates the processing of multimodal datasets; Section 4 details our proposed model framework while the experimental datasets and environment are presented in Section 5; Section 6 presents the experimental results and Section 7 concludes the paper with suggestions for future research.

2. Related work

2.1. Emotion recognition in conversations

Emotion recognition has been a hot research topic in recent decades. Because of the proliferation of conversational data and the potential application of ERC in many systems fields, ERC has begun to be widely used in the fields of cognitive science, social psychology, and natural language processing. In the existing ERC-related research, there are four main modeling methods, which are context-free, discourse context, speaker, and speaker-differentiated modeling. We present the existing research work as follows.

The context-free modeling method does not consider the context relationship between dialogues, and only uses the current dialogue information for emotion recognition. For instance, Kim et al. [9] used CNN to obtain vector representations of discursive texts to classify sentiment labels. This model, however, does not fully utilize the semantic information of the conversational context.

The discourse context modeling method mainly improves the accuracy of emotion recognition by capturing the contextual relationship of the dialogue. Poria et al. [10] used LSTM, and extracted the semantic vector representation of the discourse context, which was further used for sentiment classification. Huang et al. [11] implemented the HRLCE framework, which consists of two parts: a sentence encoder and a context encoder. The sentence encoder utilizes ELMo [12], GLOVE [13], and Deepmoji [14] to obtain vector representations of sentences, and LSTM to obtain semantic information of the discourse context. Satt et al. [15] implemented the CNN-LSTM method and applied it to the spectral domain, which obtained better performance. However, all of these methods ignore the influence of the speaker and his own dependencies on mood changes.

The speaker modeling method introduces the speaker relationship on the basis of the discourse context to further improve the accuracy of emotion recognition, which is currently the mainstream method. Hazarika et al. [16] proposed a CMN model in a speaker-based approach to modeling. The model extracts the characteristics of each speaker's discourse context separately, obtains its vector representation, and then introduces an attention mechanism to fuse the historical information of the speaker with the current discourse, thereby simulating the interaction between different speakers and the influence of the speaker's state on the current semantic information. The final output result is used to classify the emotions. However, a CMN can only simulate the interaction between two speakers. Based on CMN, Hazarika et al. [17] implemented the ICON model, which model uses GRU to connect the outputs of individual speakers in the CMN, and examines the interaction between semantic information of different speakers' historical discourses. ICON clarifies speaker-level modeling which does not use a multimodal dataset though. Lin et al. [18] modeled the historical discourse of both the current speaker and another interlocutor, as well as the discourse information of the current speaker through the IANN model. These three feature vectors were fused together with an attention mechanism to output the sentiment categories. Ghosal et al. [1] used the DialogueGCN to model the dependencies of the speaker and himself, taking into account the positional relationship between the target discourse and other discourses. However, due to the limitations of GCN, the network cannot become too deep. Otherwise, there will be over-smoothing. Zheng et al. [3] proposed DECN, a model which simulates the interaction between speakers through GGCN and corrects errors in emotion recognition strategies. Sheng et al. [6] proposed SumAggGIN, a two-stage summarization and aggregation graph reasoning network which models sentiment phrases related to the topic and dependencies on adjacent discourses in a global-to-local manner.

Although considerable progress has been made in research on ERC regarding multimodal features, existing models tend to establish an algorithmic model based on the semantic information of the conversational context and the dependency between the speaker level while ignoring the heterogeneity of the network nodes. It should be noted that the importance of different nodes shall vary.

Among the modeling approaches based on distinguishing between speakers, Majumder et al. [19] constructed the DialogueRNN model, which highlights the importance of knowing which speaker presents what information. Thus, DialogueRNN uses three different GRU networks to model speaker information, semantic information of the conversational context, and emotional information. Party GRU is used to know the speaker's status during a conversation, Global GRU is used to establish dependency between the speaker and his utterances, and Emotion GRU is used to obtain the emotion tags. These three types of information are ultimately combined to get better emotion classification. However, DialogueRNN does not incorporate speaker features into the model, which we believe is important for establishing long-term dependencies on textual contexts.

2.2. Heterogeneous graph neural network

In the past few years, Graph Neural Networks (GNN) have achieved impressive performance in various tasks due to their ability to convert graph-structured data into low-dimensional vector representations.

However, most graph data in the real world are heterogeneous, and traditional GCN models cannot handle them well. In view of this, some researchers began to pay attention to heterogeneous graphs. For example, Wang et al. [20] proposed a Heterogeneous graph Attention Network (HAN), which used a heterogeneous graph neural network with a hierarchical attention mechanism that included node-level information and semantic-level information. Based on the meta-path, the node-level attention mechanism was mainly used to learn the degree of importance between the central node and its neighbor nodes. The semantic-level attention mechanism was mainly used to learn the degree of importance between different meta-paths. HAN thoroughly considered the semantic information between nodes and meta-paths, which achieved good performance on multiple heterogeneous graph benchmark datasets. However, designing meta-paths for heterogeneous graphs with different properties requires researchers to possess certain domain knowledge. Zhang et al. [21] proposed Heterogeneous Graph Neural Network (HetGNN), which first used a random walk strategy to sample heterogeneous neighbor nodes with strong correlations between attributes for the central node and each node type grouped. Second, HetGNN performed an encoding operation on the feature vector of each group's heterogeneous attributes to obtain the embedded representation of each node. Then, HetGNN performed information aggregation on the node embedding representations of different groups to obtain node embedding representations with rich semantic information. Finally, the obtained node embedding representation was used to perform the graph node classification task. HetGNN had achieved good results on many graph data mining tasks. However, HetGNN assumed that different types of node features were in the same representation space. Fu et al. [22] proposed the Metapath Aggregated Graph Neural Network (MAGNN), which consisted of three modules: 1. Obtaining the embedded representation of nodes by transforming node attributes; 2. Performing semantic information between nodes within the meta-path polymerization. 3. Feature fusion of multiple meta-paths was performed between meta-paths. MAGNN comprehensively considered node information aggregation within meta-paths and feature fusion between meta-paths. On a large number of real heterogeneous graph benchmark datasets, MAGNN achieved more accurate classification results. However, MAGNN only used the attention mechanism to perform a weighted sum operation on different meta-paths, ignoring the dynamic interaction process of information between different meta-paths.

To tackle such problems, this paper proposes a Multivariate Message Passing Graph Convolutional Network Model (MMPGCN). Unlike current approaches based on heterogeneous graph neural networks, MMPGCN is not based on meta-paths to aggregate the semantic information of surrounding neighboring nodes but adaptively assigns a propagation weight to each node by defining the concept of homogeneity rate to achieve the aggregation of semantic information of surrounding neighboring nodes.

3. Preliminary

3.1. Problem definition

This paper focuses on the emotional changes between speakers in multiple dialogues, and n speakers participating in the dialogue are expressed as $p_1, p_2, \ldots, p_n$. U represents a set of contextual utterances spoken by n speakers in a conversation, $U = \{u_1, u_2, \ldots, u_m\}$, and m represents the number of utterances. $L = \{l_1, l_2, \ldots, l_m\}$ is the set of sentiment labels for each utterance. The set U can be represented as $U_1 \bigcup U_2 \bigcup \ldots \bigcup U_n$, where U_j represents the dialogue of speaker p_j, $j \in \{1, 2, \ldots, n\}$. To clarify the relationship between speakers and utterances, we define a set $S = \left\{ s_1^{p_i}, s_2^{p_j}, \ldots, s_m^{p_k} \mid i, j, k \in \{1, 2, \ldots, n\} \right\}$ and arrange the utterances in chronological order. Here, $s_i^{p_j} \in S$ is the ith utterance spoken by speaker p_j.

The purpose of our research is to infer the emotional state of the speakers. For an utterance $s_t^{p_j}$ at time instant t, $t \in \{1, 2, \ldots, m\}$, we need to detect the emotion of speaker p_j at instant t. Since the relationship between context and speaker needs to be modeled based on historical utterances, the historical utterance sets of speakers $p_1, p_2, \ldots, p_n$ are represented by $H_1, H_2, \ldots, H_n$ respectively. To limit the number of context and speaker relations, we employ a sliding window size for splitting historical utterances. Assuming that the size of the sliding window is K, the historical utterance formula of speaker p_j is as follows:

$$H_j = \left\{ u_i \mid i \in [t-K, t-1], u_i \in U_j, \mid H_j \mid \leq K \right\} \tag{1}$$

Among them, U_j represents the contextual utterance set of the speaker p_j. To construct reading relations between speakers, we construct a fully connected graph over the utterances belonging to the context window K. Then, we use GCN to aggregate the dialogue context information between speakers to complete the speaker's emotion prediction. Table 1 shows that when the dialogue context window K = 6, the historical dialogue context represents the multiple dialogue relations.

Table 1
Suppose there are three speakers U_a, U_b, U_c, when the dialogue context window size K = 6, the speaker's conversation as follows:

U	$u_1^a, u_2^b, u_3^a, u_4^c, u_5^b, u_6^c$
U_a, U_b, U_c	$u_1^a, u_3^a, u_2^b, u_5^b, u_4^c, u_6^c$
Test discourse	u_7^b
H_a, H_b, H_c	$u_1^a, u_3^a, u_2^b, u_5^b, u_4^c, u_6^c$

3.2. Text feature extraction

For the utterance text, this paper preprocesses the data first, uses the Tokenizer method to segment the utterance text, and generates the mapping relationship between words and the utterance text. Then input the divided words into the Roberta [7] pre-training model for fine-tuning to obtain a 100-dimensional word embedding vector, then the feature vector of the ith discourse text can be expressed as $f_i^t = \left\{ w_1, w_2, \ldots, w_k \middle| w_j \in R^{d_w} \right\}$, where k is the number of word segmentation, w_k is the feature vector of the kth word segmentation, and d_w is the word embedding dimension, $d_w = 100$. The word embedding vector contains rich semantic information, eliminates the ambiguity of words, and enables the model to approach more contextual information.

3.3. Audio feature extraction

The frequency and pitch in audio features can correctly reflect the emotional state of the speaker at the moment. In existing research on audio feature extraction, there are two mainstream research methods, which are respectively based on time-domain signals and frequency-domain signals [23]. Follow previous research work [1,24,25], the paper utilizes OpenSMILE to extract audio feature. Specifically, we input 16-bit PCM WAV format audio to IS13_ComParE1 extractor[1] for feature extraction and obtain a 6373-dimensional semantic vector. Since the feature dimension of the audio vector is too high, we also use the L2-based feature selection method to select low-dimensional feature vectors, and the audio feature in the ith utterance can be represented by $f_i^a \in R^{d_a}$, where d_a is the dimension of audio feature, and $d_a = 100$.

3.4. Visual feature extraction

For the extraction of video features, since changes in facial expressions can best reflect a speaker's emotional state, this paper proposes to use a 3-dimensional convolutional neural network (3D-CNN) [26] to perform on the speaker's facial features to enhance the model's

[1] http://audeering.com/technology/opensmile.

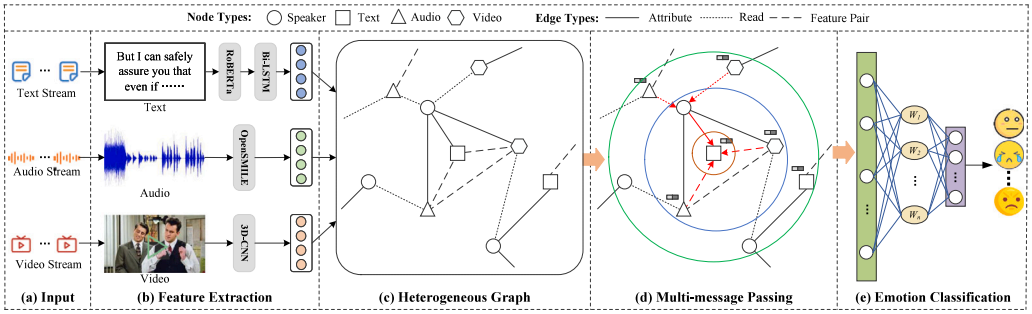


Fig. 3. MMPCGN is composed of the discourse context, video and audio features, which forms a heterogeneous graph. The information of neighbor nodes is aggregated through heterogeneous graph convolutional neural network to obtain rich semantic information.

ability to understand the emotions of the discourse. 3D-CNN is capable of extracting deep features of the human face and capturing subtle changes in facial expressions. The network consists of a convolutional layer, a pooling layer, and a fully connected layer, which inputs the video frame vectors into the network. After a series of convolution and pooling operations, the network is nonlinearly transformed by the ReLU function. Finally, the resulting feature vector passes through the fully connected layer and obtains a low-dimensional vector, which is used as a visual feature. Specifically, the visual feature vector in the ith utterance is denoted by $f_i^v \in R^{d_v}$, where d_v is the dimension of visual feature, and $d_v = 512$.

4. Methodology

As existing models overlook the heterogeneity of dialogue nodes in graph networks, this paper proposes a multivariate messaging framework to aggregate video features, audio features, and discourse features. The framework consists of feature extraction, heterogeneous graph construction, multi-message passing, and sentiment classification. In feature extraction, the text, audio, and visual information in each utterance are encoded into feature vectors with rich semantics through the Roberta pre-training model, OpenSMILE model, and 3D-CNN model, and then the text feature vectors are input to the Bi-LSTM model to obtain rich contextual semantic information. In heterogeneous graph construction, speaker, text, audio, and video features are used as nodes in the graph network, and dialogue, reading, and feature-pair relations are used as edges. All this helps construct a speaker-level dependency. In multi-message passing, considering the heterogeneity of given different nodes, this paper defines the isomorphism rate of network nodes based on the heterogeneous graph, and dynamically assigns weights to each edge of the graph node according to the isomorphism rate. To improve the generalization ability of the model, this paper allows all neighbors to share weights and reduces the number of network parameters, and thus aggregates node information and obtains node representations with rich semantic features. Finally, the obtained semantic information is entered into the classifier to process sentiment classification with the maximum probability. The model framework is shown in Fig. 3.

4.1. Sequential context information modeling

Since the utterance is actually a list of words arranged in a certain order sequence, its context and semantic information will be transmitted in this order accordingly. Bi-LSTM is obtained by splicing LSTMs in forward order and reversed order, which can better capture bidirectional contextual semantic information. Therefore, in this paper, we use Bi-LSTM to model the discourse context, input f_i^t into the Bi-LSTM neural network, and extract the contextual feature representation f_i^w with rich semantic information.

4.2. Speaker interaction modeling with multiple message passing

This study builds a multivariate message-passing framework to handle the heterogeneity of nodes in graph networks. As shown in Fig. 4, we have constructed two message aggregation methods: one is to aggregate node information with a feature pair relationship, and the other is to aggregate node information with dialogue and reading relationships. The graph convolution operation through our designed multivariate message aggregation mechanism can enhance the feature representation ability of nodes. Furthermore, based on this framework, the speaker-level context is modeled to simulate the interaction between speakers, generating highly abstract features containing the speaker-level semantic context.

A graph $G = (V, E, R, W)$ is constructed, which has a set of nodes $V = (v_1, v_2, \ldots, v_N)$, a set of graphs of edges E, and the relationship between nodes R. And the weight between edges W, where N is the number of nodes in the graph network. The speaker's video features, audio features, and utterance features can be considered as nodes in the graph. The graph is further used to model the speaker-level context and simulate the interaction between speakers.

In order to solve the problem of node heterogeneity in the graph network, this paper proposes the concept of node homogeneity, which measures the difference of nodes by calculating the similarity between nodes with message-passing relationships and assigns weights to each edge in the graph network based on the node homogeneity rate. Through aggregating information about the neighbors around the nodes, rich semantic information can be obtained in the end. The calculation formula of the node homogeneity rate is as follows:

$$\alpha_v = \frac{sim(u, v)}{\sum_{u \in P(v)} sim(v, x) + \sum_{x \in D(v)} sim(v, x)} \tag{2}$$

Among them, α_v represents the homogeneity rate of node v, $P(v)$ is the set of neighbor nodes that have a feature pair relationship with node v, $D(v)$ is the set of neighbor nodes that have dialogue and reading relationships with node v, and $sim()$ is a similarity function. In addition, when the similarity between node v and feature pair nodes is larger, the homogeneity rate of node v is higher, and vice versa.

For edge weights, in order to realize the difference between feature vectors of different modes in the information aggregation, this paper sets corresponding weights for each edge based on the homogeneity rate, and the softmax function is used to ensure that the sum of the weights of each edge is one. The edge weight between two nodes u and v is defined as:

$$\delta_{u,v}^t = softmax\left(\frac{\exp(\alpha_v)}{\sum_{k \in \mathcal{N}_u} \exp(\alpha_k)}\right) \tag{3}$$

Among them, t represents the tth time in a period of time, $\mathcal{N}_u$ is the neighborhood nodes of node u in the graph.

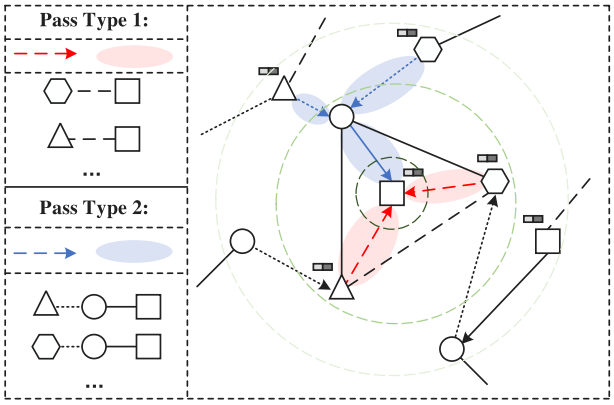


Fig. 4. The multivariate messaging schematic. We aggregate two kinds of information around a node, one is to aggregate node information with feature pair relationship, and the other is to aggregate node information with dialogue and reading relationship.

To aggregate the information about surrounding neighbor nodes and further pass the information about discourse context in the network, we adopt the following formula:

$$h_v^k = \sigma\left(\left(c_{vv}^k \odot h_v^{k-1} + \sum_{u\in N(v)} \delta_{u,v}^{k-1} c_{uv}^k \odot h_u^{k-1}\right) W^k\right) \tag{4}$$

where W_k is the parameter learned by the network itself, and σ is the activation function. Here, the ReLU function is chosen as the activation function, $\odot$ represents the product of the corresponding elements, and $N(v) = P(v) + D(v)$. c_{uv} denotes a weight, a self-attention mechanism in the current study. The formula is as follows:

$$c_{uv}^k = \tanh\left(\left[h_v^{k-1} \parallel h_u^{k-1}\right] W^k\right) \tag{5}$$

To reduce the number of network parameters and improve the generalization ability of the model, this paper proposes an algorithm mechanism that assumes that all neighbor nodes share weights and transmit semantic information with the same weights. In other words, the contribution of all nodes is equal. The formula can be presented as follows:

$$h_v^k = \sigma\left(\left(c_v^k \odot h_v^{k-1} + \tilde{c}_v^k \odot \sum_{u\in N(v)} \delta_f^{k-1} h_u^{k-1}\right) W^k\right) \tag{6}$$

Meanwhile, Eq. (7) is as follows:

$$c_v^k = \tanh\left(\left[h_v^{k-1} \parallel \tilde{h}_v^{k-1}\right] W_c^k\right) \tag{7}$$

4.2.1. Emotion classification

Once the multi-message transfer framework aggregates the rich semantic information about the nodes in the graph network, the sentiment classifier splits the text, audio, and video features together. Since different modalities have different effects on emotion classification, this paper uses the self-attention mechanism to dynamically fuse multi-modal features so as to improve the effect of feature fusion and obtain new discourse features. Then, the obtained speech feature representation is input into the fully connected layer, which is subsequentially passed through the softmax activation function to obtain the corresponding emotional label probability distribution. Finally, the emotional label $\hat{y}_i$ with the maximum probability is obtained using the argmax function (see Fig. 4).

As shown in Eq. (8), the speech vector g_i, audio feature vector u_i, and video feature vector τ_i, which are rich in semantic information and obtained through graph neural network aggregation are connected.

$$h_i = concat\left[g_i, u_i, \tau_i\right] \tag{8}$$

Table 2
The division of training set, test set, and validation set on the IEMOCAP and MELD datasets, as well as the number of emotional categories and evaluation metrics. Acc = Accuracy.

Datasets	Utterance count			Dialogue count			Classes	Evaluation
	Train	Validation	Test	Train	Validation	Test		metrics
IEMOCAP	5320	490	1623	108	12	31	6	Accuracy/f1
MELD	9989	1109	2610	1038	114	280	7	Accuracy/f1

Eqs. (9) and (10) show that the self-attention mechanism is adopted to obtain the final speech representation $\hat{h}_i$ from the feature vectors obtained by connection.

$$\beta_i = softmax\left(\left[h_1, h_2, \ldots, h_N\right]^T W_\beta \left[h_1, h_2, \ldots, h_N\right]\right) \tag{9}$$

$$\tilde{h}_t = \beta_i \left[h_1, h_2, \ldots, h_N\right]^T \tag{10}$$

The final speech feature representation $\hat{h}_i$ obtained by the self-attention mechanism is input into the fully connected layer and then nonlinearly activated through the ReLU activation function to obtain feature information of the hidden layers ρ_i. The formula is defined as follows:

$$\rho_i = ReLU\left(W_l \tilde{h}_l + b_l\right) \tag{11}$$

Then, the utterance features are input into the softmax function to obtain the probability distribution of the emotion P_i corresponding to the utterance feature. The formula is as follows:

$$P_i = softmax\left(W\rho_i + b\right) \tag{12}$$

Finally, the probability distribution of emotional labels P_i is passed through the argmax function to obtain the emotional label $\hat{y}_i$ with the maximum probability corresponding to the discourse feature. The formula is specified as follows:

$$\hat{y}_i = \arg\max_t \left(P_i[t]\right) \tag{13}$$

5. Experimental setting

5.1. Implementation details

All the research mentioned thus far is conducted on two NVIDIA Tesla P4 servers with a total memory capacity of 16G. We use Python3.7 as our programming language, and the deep learning framework is Pytorch1.8.1. Adam [27] is adopted as our optimization algorithm, wherein the batch size is set to 32, the number of iterations is 60, the initial learning rate is 3e−4, and the L2 weight attenuation coefficient is attenuated to 1e−5.

5.2. Datasets used

Experiments have been conducted on two benchmark datasets of different sizes for ERC, IEMOCAP [28] and MELD [29] to test our algorithmic model. The evaluation indicators for the validation set, training set, and test set of the benchmark datasets, as well as for model effectiveness, are shown in the Table 2:

IEMOCAP: The IEMOCAP dataset contains videos of binary conversations involving five men and five women. These videos comprise five stages of the conversations, and during each stage, a binary conversation is assigned between a man and a woman. Each utterance is annotated with an emotional label (happy, neutral, sad, angry, frustrated, and excited). In our study, the dialogues during the first four stages are chosen as training sets and validation sets, while the dialogues in the last stage is adopted as test sets.

MELD: The MELD dataset is a multi-party conversation dataset that contains more than 13,000 utterance texts. Each conversation stage involves three or more speakers, and MELD is a multimodal dataset, and each text is labeled with a particular emotion (anger, sadness, disgust, joy, surprise, neutrality, and fear).

5.3. Baselines and state of the art

We compare our model with the following baseline models:

CNN: The CNN proposed by Kim et al. [9] is a baseline model for the classification of conversational texts. As it does not model the utterance context or the dependencies between speakers. CNN cannot use multimodal data.

Bc-LSTM: The bidirectional LSTM proposed by Poria et al. [10] et al. captures the semantic information of the discourse context from the speaker's historical context discourse and its current discourse. It, however, ignore the positional relationship between the speaker and the conversational context, as well as the interaction between speakers.

CMN: CMN proposed by Hazarika et al. [12] uses GRU to obtain rich contextual, semantic vector representations and inputs them into a memory network, realizing modeling of long-term contextual information. However, this model can only detect emotional changes between two speakers.

DialogueRNN: Majumder et al. [10] proposed DialogueRNN, which is a recurrent neural network using three different GRU networks to model speaker information, semantic information in the context of the conversation, and emotional information. Meanwhile, a self-attention mechanism is introduced to the network to obtain attention scores from rich semantic vectors, which fully considers the relationship between the dialogue context and the speaker. The model can be extended to multimodal datasets.

AGHMN: Jiao et al. [30] proposed AGHN, a gated hierarchical memory network based on attention mechanisms. The model introduces a self-attention mechanism into the GRU for determine the weights of the recent conversation context vector and the conversation context vector from distant memory. It is capable of identifying emotions in real time.

DialogueGCN: Ghosal et al. [1] proposed DialogGCN, which utilizes GCN to assess the speaker's ability to influence the model's understanding of contextual semantic information. The model combines discourse context semantic encoding with speaker-level information encoded to improve model performance, which can be applied to multivariate conversations relationships.

SumAggGIN: This is the most advanced model for ERC tasks so far. Proposed by Sheng et al. [6], SumAggGIN consists of a two-stage summarization and aggregation graph inference network. The network is able to detect subtle differences between with phrase-level utterances.

DialogueCRN: DialogueCRN [31] proposes an emotional cue-aware emotion recognition model, which extracts contextual emotional cues by building a multi-round reasoning module.

DisGCN: DisGCN [32] proposes a discourse-aware graph neural network to model the importance of discourse structure for context information and speaker information exploration. It uses GCN and gated convolution to extract speaker information and discourse structure information, respectively.

6. Results and discussion

6.1. Comparison with state-of-the-art and baseline methods

Our proposed model MMPGCN is compared with the listed baseline models and the current state-of-the-art model SumAggGIN and DisGCN. The experimental results show that our proposed model has outperformed existing models on two benchmark datasets.

On the IEMOCAP dataset, the accuracy of our model is 68.9%, 1.9% higher than SumAggGIN, and 4.5% higher than DisGCN; the f1 value is 68.0%, 1.3% higher than SumAggGIN, and 4.5% higher than DisGCN. On the ME-LD dataset, the accuracy of our model is 60.7%, 2.2% higher than SumAggGIN, and 1.6% higher than DisGCN; the f1 value is 59.3%, 2.7% higher than Sum-AggGIN, and 3.1% higher than DisGCN. The accuracy and f1 value of each label in the IEMOCAP and MELD datasets obtained by our proposed model and other baseline models are shown in Tables 3 and 4, respectively.

Compared with other baseline models, MMPGCN has better performance improvement on the two chosen benchmark datasets. We believe that the main reason lies in the difference of model. At present, mainstream methods primarily include discourse context-based modeling and modeling based on speaker-levels. The current state-of-the-art method, SunAggGIN, comprehensively examines the relationship between speaker-level features and phrase-level features and uses a two-stage GNN to aggregate rich semantic information, which enhances the performance to a certain extent. DisGCN uses GCN and gated convolution to extract speaker information and discourse structure information, respectively. However, these baseline models ignore the heterogeneity between multimodal nodes, which, we think, has an important impact on the model's understanding of the emotional changes in discourses.

6.2. Error analysis

We also analyze the labels predicted by the model. As shown in Fig. 5(a), the confusion matrix used on the IEMOCAP dataset found that our model incorrectly classified the "happy" emotion as "excited", while the "frustrated" emotion is labeled as "neutral". This may be caused by the slight differences between the two labels. By increasing the dataset size, we think we can make the model note the subtle differences between the two and obtain more accurate results.

As shown in Fig. 5(b), the confusion matrix for the MELD dataset observed found that the model misclassified the "neutral" as "surprise", "sadness", "joy", and "anger", which are included in the IEMOCAP dataset. Likewise, the difference between these labels and the "neutral" emotion is small, so the model fails to tell the difference. In the "fear" and "disgust" labels, the prediction accuracy of our model is 7.7% and 9.1%, respectively, and the f1 value is 3.2% and 2.6%, respectively. Through analysis, it is found that the dataset sizes in these two labels are 50 and 68, respectively. It is assumed that the two datasets are too small, which cannot provide sufficient useful information for the model. Instead, the model can only use the semantic information in other labels to proceed with the classification, which results in very low accuracy and f1 value. In future studies, we hope to improve the performance of the model by increasing the dataset size.

6.3. Ablation study

Our proposed model MMPGCN is innovative in that it introduces the Roberta pre-training model, uses Bi-LSTM to model contextual semantic information, comprehensively considers the heterogeneity between different modalities, and proposes the concept of homogeneity rate. Based on the concept, the model is able to distinguish between influences from different modalities, thus dynamically assigning edge weights to graph nodes. In addition, in order to ensure that the model can converge effectively, we adopt the method of sharing parameters for the edge weights of neighbor nodes in the graph structure. To verify the effectiveness of our model, we perform ablation experiments on the IEMOCAP dataset, removing one module at a time to determine the respective contribution to the model.

The results are shown in Table 5. If MMPGCN is not considered in the model to model the interaction between speakers, the f1 value of the model will drop by 10.1%. It is believed that taking into account the heterogeneity between different modalities could help the model to give more weight to the more influential modal nodes, thereby improving the performance of the model. In addition, it is also necessary to model the interaction between speakers, which will help the model understand how past utterances affect the emotional changes of future utterances. While using MMPGCN to model the interaction between speakers, without parameter sharing of the edge weights of surrounding neighbor nodes, the f1 value of the model will drop by 5.7%. We believe this is because the complexity of the model is too high, which prevents

Table 3
Comparison with other baseline models on the IEMOCAP dataset, Acc. = Accuracy, Average(w) = Weighted average.

Methods	IEMPCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average(w)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN[a]	27.7	29.8	57.1	53.8	34.3	40.1	61.1	52.4	46.1	50.0	62.9	55.7	48.9	48.1
CNN	24.4	32.3	55.9	57.9	51.8	44.0	29.2	43.4	68.0	34.1	57.4	40.0	42.6	42.2
bc-LSTM	29.1	34.4	57.1	60.8	54.1	51.8	57.0	56.7	51.1	57.9	67.1	58.9	55.2	54.9
CMN	25.0	30.3	55.9	62.4	52.8	52.3	61.7	59.8	55.5	60.2	**71.1**	60.6	56.5	56.1
DialogueRNN	25.6	33.1	75.1	78.8	58.5	59.2	64.7	**65.2**	**80.2**	71.8	61.1	58.9	63.4	62.7
DialogueGCN	40.6	42.7	**89.1**	**84.4**	61.9	63.5	67.5	64.1	65.4	63.0	64.1	66.9	65.2	64.1
DialogueCRN	**71.4**	51.9	75.8	78.2	66.1	59.8	**78.5**	64.1	68.9	77.7	54.9	60.1	66.4	65.7
SumAggGIN	56.7	54.2	86.8	79.1	62.9	65.3	64.6	62.2	76.2	78.4	63.4	61.6	66.8	66.7
DisGCN	71.1	**56.9**	68.6	76.4	66.6	57.4	74.2	54.3	74.5	76.4	51.1	59.2	64.4	63.8
MMPGCN	61.3	56.7	84.3	82.5	**72.3**	**65.8**	56.3	54.3	71.5	**78.8**	64.6	**69.9**	**68.9.**	**68.0**

[a] Indicates that only text modal data is used. Otherwise, it indicates that three modal data of text, video, and audio are used.

Table 4
Comparison with other baseline models on the MELD dataset, Acc. = Accuracy, Average(w) = Weighted average.

Methods	MELD															
	Neutral		Surprise		Fear		Sadness		Joy		Disgust		Anger		Average(w)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN[a]	76.2	74.9	43.3	45.5	4.6	3.7	18.2	21.1	46.1	49.4	8.9	8.3	35.3	34.5	56.3	55.0
CNN	77.7	76.0	48.3	42.4	2.5	4.1	19.2	20.4	48.2	49.3	5.4	**5.9**	36.0	31.8	57.2	55.4
bc-LSTM	**78.4**	73.8	46.8	47.7	3.8	**5.4**	22.4	25.1	51.6	51.3	4.3	5.2	36.7	38.4	57.5	55.9
DialogueRNN	72.1	73.5	**54.4**	49.4	1.6	1.2	23.9	23.8	**52.0**	**50.7**	1.5	1.7	41.0	41.5	56.1	55.9
DialogueCRN	70.9	75.7	47.3	47.1	0.0	0.0	**34.0**	13.2	41.9	49.7	0.0	0.0	41.6	35.6	58.3	54.9
SumAggGIN	–	–	–	–	–	–	–	–	–	–	–	–	–	–	58.5	56.6
DisGCN	70.8	76.6	42.7	46.1	1.1	1.5	32.0	16.9	50.3	50.1	2.3	1.9	38.2	39.9	59.1	56.2
MMPGCN	77.7	**78.6**	53.5	**53.8**	**7.7**	3.2	32.3	**25.2**	50.7	53.3	**9.1**	2.6	**43.1**	**45.0**	**60.7**	**59.3**

[a] Indicates that only text modal data is used. Otherwise, it indicates that three modal data of text, video, and audio are used.

Table 5
Ablation experiment of MMPGCN model on the IEMOCAP dataset.

Roberta	Bi-LSTM	Shared parameters	MMPGCN	F1
+	+	–	–	57.9
+	+	–	+	62.3
+	–	+	+	66.6
–	+	+	+	62.1
+	+	+	+	68.0

Table 6
Exploring the scalability of MMPGCN on the IEMOCAP dataset. Att = Attention.

Method	F1
GCN	58.4
MMPGCN	60.6
GCN+Att	64.1
MMPGCN+Att	68.0

Table 7
Experimental results of different modal features on IEMOCAP and MELD datasets. T, A, and V represent text, audio, and video features, respectively.

Modality	IEMOCAP	MELD
	F1	F1
T	64.3	55.1
T+V	64.8	56.0
T+A	66.4	56.6
T+A+V	67.2	57.2

the model from converging efficiently. Without considering the use of Bi-LSTM to model sequential contextual semantic information, the f1 value of the model will be 1.4%. We speculate that this is because the current utterance may be closely related to the contextual utterance. If only the current utterance is used to judge the emotion contained in a sentence, it would not be very objective. Obtaining rich semantic information through the Roberta pre-training model benefits the model proposed in this paper to enhance the ability to understand emotional labels. If this part is removed, the effect of the model will drop by 2.9%. We believe that word vectors with rich semantic information help the model understand the subtle differences between labels.

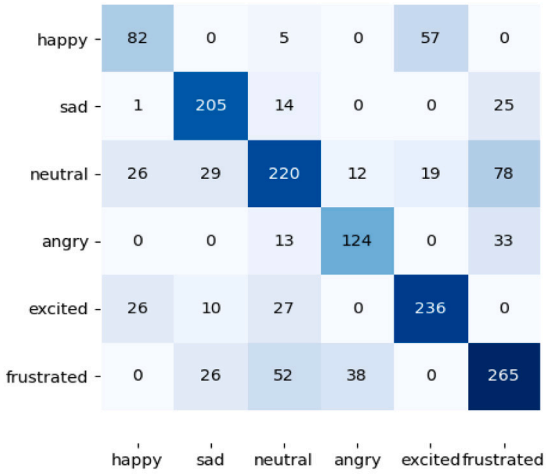
In addition, to explore the scalability of the MMPGCN model proposed in this paper, we conduct comparative experiments on the IEMOCAP benchmark dataset. As shown in Table 6, without using the multivariate message passing mechanism proposed in this paper but only using ordinary GCN and GCN+Att, the f1 values of the model under the IEMOCAP benchmark dataset are 58.4% and 64.1%, respectively. 2.2% and 3.9% lower than MMPGCN and MMPGCN+Att. Therefore, it is believed that the message-passing mechanism proposed in this paper can be transferred to other graph neural networks.

6.4. Importance of multimodal features

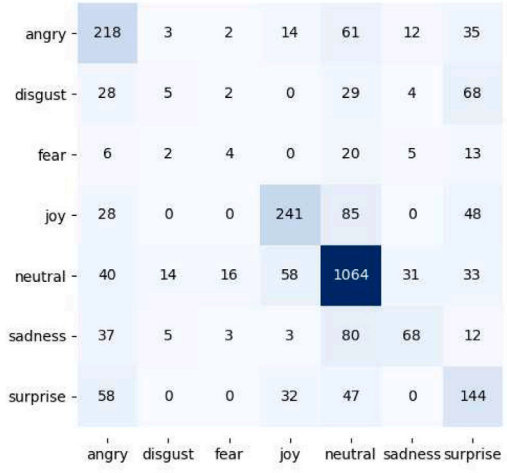
As shown in Table 7, when using only text features, our model MMPGCN achieves an average F1 value of 64.3% and 55.1% on the IEMOCAP and MELD datasets, respectively. When using text and video features, the emotion recognition effect of the model can be improved, with an average F1 value of 64.8% and 56.0%, respectively. When text and audio features are used, the emotion recognition effect of the model is better than using text and audio features. When text, video, and audio features were used, the model performed best in emotion recognition, with average F1 values of 67.2% and 57.2%, respectively. Experimental results illustrate the necessity of multimodal emotion recognition.

7. Conclusion

This paper has proposed a heterogeneous multi-dimensional messaging framework based on multimodality for detecting emotions in discourse. Compared with current mainstream methods, our proposed method considers the heterogeneity between different modes and provides richer feature information for graph networks. Compared to the current research work, the model designed in this paper has achieved some performance improvement on two different datasets for ERC. In



(a) Confusion matrix obtained by MMPGCN on the IEMOCAP dataset.



(b) Confusion matrix obtained by MMPGCN on the MELD dataset.

Fig. 5. The MMPGCN model predicts the sentiment labels of the IEMOCAP dataset and the MELD dataset.

the future, we plan to explore different types of nodes and edges in multimodal graph networks at the node level and semantic level, which will aid us in obtaining richer semantic information and optimizing the model.

CRediT authorship contribution statement

Tao Meng: Investigation, Conception and design of study, Acquisition of data, Software, Writing the original manuscript. **Yuntao Shou:** Methodology, Analysis and interpretation of results, Writing – review & editing. **Wei Ai:** Funding acquisition, Resources, Reviewing and editing. **Jiayi Du:** Validation, Supervision. **Haiyan Liu:** Software. **Keqin Li:** Reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors deepest gratitude goes to the anonymous reviewers and AE for their careful work and thoughtful suggestions that have helped improve this paper substantially. This work is supported by National Natural Science Foundation of China (Grant No. 69189338), Research Foundation of Education Bureau of Hunan Province of China (Grant No. 20B625, No. 22B0275), Changsha Natural Science Foundation, China (Grant No. kq2202294), and program of Research on Local Community Structure Detection Algorithms in Complex Networks, China (Grant No. 2020YJ009).

References

[1] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, 2019, arXiv preprint arXiv:1908.11540.
[2] T. Ishiwatari, Y. Yasuda, T. Miyazaki, J. Goto, Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7360–7370.
[3] Z. Lian, B. Liu, J. Tao, DECN: Dialogical emotion correction network for conversational emotion recognition, Neurocomputing 454 (2021) 483–495.
[4] C. Shi, Y. Li, J. Zhang, Y. Sun, S.Y. Philip, A survey of heterogeneous information network analysis, IEEE Trans. Knowl. Data Eng. 29 (1) (2016) 17–37.
[5] L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 7370–7377.
[6] D. Sheng, D. Wang, Y. Shen, H. Zheng, H. Liu, Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4153–4163.
[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[8] L. Yang, M. Li, L. Liu, C. Wang, X. Cao, Y. Guo, et al., Diverse message passing for attribute with heterophily, in: Thirty-Fifth Conference on Neural Information Processing Systems, 2021.
[9] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751.
[10] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 873–883.
[11] C. Huang, A. Trabelsi, O.R. Zaïane, Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert, 2019, arXiv preprint arXiv:1904.00132.
[12] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, 2018, arXiv:1802.05365.
[13] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[14] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, S. Lehmann, Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, 2017, arXiv preprint arXiv:1708.00524.
[15] A. Satt, S. Rozenberg, R. Hoory, Efficient emotion recognition from speech using deep learning on spectrograms., in: Interspeech, 2017, pp. 1089–1093.
[16] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the Conference. Association for Computational Linguistics. North American Chapter. Meeting, Vol. 2018, NIH Public Access, 2018, p. 2122.
[17] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, Icon: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2594–2604.
[18] S.-L. Yeh, Y.-S. Lin, C.-C. Lee, An interaction-aware attention network for speech emotion recognition in spoken dialogs, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6685–6689.

[19] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6818–6825.
[20] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, P.S. Yu, Heterogeneous graph attention network, in: The World Wide Web Conference, 2019, pp. 2022–2032.
[21] C. Zhang, D. Song, C. Huang, A. Swami, N.V. Chawla, Heterogeneous graph neural network, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 793–803.
[22] X. Fu, J. Zhang, Z. Meng, I. King, Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding, in: Proceedings of the Web Conference 2020, 2020, pp. 2331–2341.
[23] Y. Luo, Z. Chen, N. Mesgarani, Speaker-independent speech separation with deep attractor network, IEEE/ACM Trans. Audio, Speech, Lang. Process. 26 (4) (2018) 787–796.
[24] Z. Li, F. Tang, M. Zhao, Y. Zhu, EmoCaps: Emotion capsule based model for conversational emotion recognition, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 1610–1618.
[25] Z. Lian, B. Liu, J. Tao, Ctnet: Conversational transformer network for emotion recognition, IEEE/ACM Trans. Audio, Speech, Lang. Process. 29 (2021) 985–1000.
[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[27] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[28] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Res. Eval. 42 (4) (2008) 335–359.
[29] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2018, arXiv preprint arXiv:1810.02508.
[30] W. Jiao, M. Lyu, I. King, Real-time emotion recognition via attention gated hierarchical memory network, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 8002–8009.
[31] D. Hu, L. Wei, X. Huai, Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations, in: 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, pp. 7042–7052.
[32] Y. Sun, N. Yu, G. Fu, A discourse-aware graph neural network for emotion recognition in multi-party conversation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2949–2958.

Tao Meng received the Ph.D. degree in the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. His research interests include data mining, artificial intelligence, machine learning, natural language processing, graph and network analysis. (Email: mengtao@hnu.edu.cn)

Yuntao Shou received the B.S. degree in School of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, China, in 2023. He is currently pursuing the graduation degree with Xi'an Jiaotong University, Xian, China. His research interests include emotion recognition and graph representation learning. (Email: yuntaoshou@csuft.edu.cn)

Wei AI received the Ph.D. degree in the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. Her research interests include date mining, big data, cloud computing, and parallel computing. (Email: aiwei@hnu.edu.cn)

Jiayi Du received his Ph.D., M.Sc. and B.Sc. in computer science from Hunan University, China, in 2015, 2010 and 2004. He is currently an assistant professor in Central South University of Forest and Technology, China. His research interest includes modeling and scheduling for parallel and distributed computing systems, embedded system computing, cloud computing, parallel system reliability, and parallel algorithms. (Email: dujiayi@csuft.edu.cn)

Haiyan Liu holds an associate professor at Changsha Medical University and deputy director of the Hunan Provincial University Key Laboratory of the Fundamental and Clinical Research on Functional Nucleic Acid. Her primary research interests encompass machine learning and biological information computing. (Email: liuhy_csmu@163.com)

Keqin Li is a SUNY Distinguished Professor of Computer Science with the State University of New York. He is also a National Distinguished Professor with Hunan University, China. His current research interests include cloud computing, fog computing and mobile edge computing, energy-efficient computing and communication, embedded systems and cyber–physical systems, heterogeneous computing systems, big data computing, high-performance computing, CPU–GPU hybrid and cooperative computing, computer architectures and systems, computer networking, machine learning, intelligent and soft computing. He has authored or coauthored over 890 journal articles, book chapters, and refereed conference papers, and has received several best paper awards. He holds nearly 70 patents announced or authorized by the Chinese National Intellectual Property Administration. He is among the world's top 5 most influential scientists in parallel and distributed computing in terms of both single-year impact and career-long impact based on a composite indicator of Scopus citation database. He has chaired many international conferences. He is currently an associate editor of the ACM Computing Surveys and the CCF Transactions on High Performance Computing. He has served on the editorial boards of the IEEE Transactions on Parallel and Distributed Systems, the IEEE Transactions on Computers, the IEEE Transactions on Cloud Computing, the IEEE Transactions on Services Computing, and the IEEE Transactions on Sustainable Computing. He is an AAAS Fellow, an IEEE Fellow, and an AAIA Fellow. He is also a Member of Academia Europaea (Academician of the Academy of Europe). (Email: lik@newpaltz.edu)