Bi-stream Graph Learning based Multimodal Fusion for Emotion Recognition in Conversation

Nannan Lu[a], Zhiyuan Han[a], Min Han[b], Jiansheng Qian[a,*]

[a]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, 221000, China
[b]School of Control Science and Engineering, Dalian University of Technology, Dalian, 116000, China

Abstract

Emotion Recognition in Conversation (ERC) is the process of automatically detecting and understanding emotions expressed in a conversation, which plays an important role in human-computer interaction. A conversation generates different modality data including words, tone of voice and facial expression. Multimodal ERC can fuse the information from multiple views to comprehensively model emotion dynamics in a conversation. Graph Neural Networks (GNNs) are employed by multimodal ERC to learn intra-modal long-range contextual information and inter-modal interaction. However, fusing different modalities within a graph may generate the conflict of multimodal information and suffer from data heterogeneity issue. In the paper, we propose a novel Bi-stream Graph Learning based Multimodal Fusion (BiGMF) approach for ERC. It consists of a unimodal stream graph learning for modeling the intra-modal long-range context information and a cross-modal stream graph learning for modeling the inter-modal interactions, which uses GNNs to learn the intra- and inter-modal information in parallel. The separation learning scheme can successfully alleviate the conflict and heterogeneity in multimodal data fusion, and promote the explicitly modeling of cross-modal relations. The experimental results on two public datasets further verify that the superiority of the proposed approach compared to the SOTA approaches.

Keywords: Emotion Recognition in Conversation, Multimodal Fusion, Graph Neural Networks, Contextual Information, Inter-modal Interaction

1. Introduction

Emotion is an external expression of human's internal psychological states, which influences human's decisions and behaviors. In human-computer interaction, the ability of machines to perceive and recognize human emotions can facilitate emotional communication, enhance the machines understanding of users' emotional needs, and enable a more intelligent and personalized interaction process. Emotion Recognition in Conversation (ERC) is a significant research direction in emotion recognition aiming at detecting the emotion states of speakers during the conversation, which is crucial for the advancement of human-computer interaction. With the rapid growth of artificial intelligence, ERC's applications are expanding to dialogue generation [1], recommender system [2], opinion mining [3], and medical diagnosis [4].

In a dialogue, emotions flow between the context, meaning that the current emotion is affected by the context. Therefore, many existing ERC methods focus on modeling the contextual information. Jiao et el. [5] proposed a hierarchical Gated Recurrent Units (GRUs) framework, including a lower-level GRU for learning the individual utterance embeddings and an upper-level GRU for capturing the context of utterance-level embeddings. COSMIC [6] modeled various aspects of commonsense knowledge, such as mental states, events, actions and cause-effect relations, to alleviate issues of detecting shifts in emotion and differentiating between related emotion classes existing in the context propagation. DialogueCRN [7] explored the cognitive factors for ERC for the first time, which imitated human's unique cognitive thinking to understand the conversational context from the cognitive perspective. DialogueGCN [8] constructed a directed graph assisted by self- and inter-speaker dependency to model each conversation, which improved the long-range context propagation. Although the above models have captured the contextual relationships in ERC to some extent, they are all based on a single data source. In fact, in a conversation, language, tone of voice and facial expressions are all expressions of emotions [9, 10, 11]. Therefore, it is necessary to study the use of multimodal information to model the contextual relationships. As an early approach, BC-LSTM [12] used three bidirectional Long Short-Term Memory (LSTM) networks to capture the contextual information from textual, acoustic and visual modalities, and then used an additional bidirectional LSTM to perform multimodal information fusion. DialogueRNN [13] used three GRUs to model the Global State, Party State, and Emotion Representation, respectively. However, the aforementioned approaches fail to capture the long-range contextual dependencies which is critical to understand the emotion dynamics by the multimodal approaches in a con-

*Corresponding author
Email addresses: lnn_921@126.com (Nannan Lu), hzy0310@163.cn (Zhiyuan Han), minhan@dlut.edu.cn (Min Han), qian_cumt@126.com (Jiansheng Qian)

versation. Thereupon, some studies leverage Graph Neural Networks (GNNs) to model the long-range contextual dependency for the multimodal ERC. ConGCN [14] employed a heterogeneous graph to model the context-sensitive and speaker-sensitive dependencies, where the nodes in the graph represented either each utterance or each speaker of the whole conversation corpus and the edges connected utterances to other utterances as well as to speakers. COGMEN [15] believed that the long-range contextual information existed in the emotion dynamics of the self- and inter-speaker, so that it modeled the self- and inter-speaker dependencies which were represented by the directed edges in the graph, and used the Relational GCN with a GraphTransformer to capture the dependencies. As to the manner of the multimodal fusion, the above approaches simply concatenate the multimodal features in a direct manner, which do not interact between different modalities. Therefore, except for the long-range contextual information, the multimodal interaction seems much more important for multimodal fusion, and needs to be further explored.

As is well known, GNNs are proficient in modeling various complex relationships for its structure advantage. Several approaches have started to employ GNNs to model the multimodal interaction in ERC. Hu et al. [16] constructed a fully connected graph with a large-scale size for ERC to learn the multimodal interaction between different modalities as well as the contextual information. In the graph, the nodes represented the features of different modalities learned from different utterances, and two types of edges were used to represent inter-modal relationships and contextual relationships, respectively, in which the former relationships were represented by the edges between the nodes belonging to different modalities from the same utterance, and the latter ones were captured by the edges between any pair of nodes belonging to the same modality. Then, the named MMGCN leveraged the spectral graph convolution to capture the two kinds of relationships in the same graph. Although the approach has achieved the dual complex relation learning via the graph-based fusion, it simply and equally treated the heterogeneous nodes of different modalities and the homogeneous nodes of same modality, and tried to use a homogeneous graph to fuse the heterogeneous information, which ignored the heterogeneity between different modalities [17]. GraphMFT [18] constructed a spatial convolution based graph attention network to fuse different modalities. It believed that the introduction of too many modalities into graph learning might bring difficulty in fusing, so that the approach reduced the modalities to be fused and then constructed three heterogeneous graphs where each graph fused two modalities. Thus, GraphMFT improved the graph construction of MMGCN. GraphCFC [19] proposed a Pair-wise Cross-modal Complementary (PairCC) strategy to achieve the complementation and fusion of different modalities. PairCC was implemented via a graph network with a GATMLP layer which could capture the intra-modal contextual and inter-modal interaction information through the edges of the graph. The construction of graphs is critical for multimodal fusion. The inappropriate graph construction may result into the conflict between the contextual information and the inter-modal interaction information. In GraphMFT and GraphCFC, the neighborhoods of any node in the graph may contain different modality information. The neighborhood nodes having the same modality as the central node propagate the intra-modal contextual information. While the others having different modality as the central node propagate the inter-modal interaction information. Even though the propagated information of the neighborhoods conflicts with each other, the approaches still fuse them, which will result into the confusion during the multimodal fusing. GA2MIF [20] used the cross-modal attention mechanism to measure the cross-modal interaction in pairs, which was achieved based on the intra-modal local and long-range contextual information captured by multi-head directed graph attention networks. The features learned after the cross-modal interaction were concatenated to ultimately complete the multimodal fusion. Nevertheless, GA2MIF calculated the correlation weights between the given modality and other modalities, and subsequently multiplied the weights and the features of the given modality, which could not realize the direct interaction between different modalities on the feature level. Therefore, the process might weaken the effect of cross-modal interaction information to the fusion, which was believed as an implicit cross-modal relation modelling way.

In order to simultaneously capture the intra-modal contextual information and the inter-modal interaction information, we propose a bi-stream graph learning based multimodal fusion (BiGMF) approach in the paper. The BiGMF consists of the unimodal stream graph learning and the cross-modal stream graph learning, which use the unimodal graph attention networks (UMGATs) and cross-modal graph attention networks (CMGATs) to capture the intra-modal contextual information and the inter-modal interaction information in an explicit way. Each UMGAT constructs a homogeneous graph to model the intra-modal contextual relationships. It uses the self-attention mechanism to replace the traditional graph attention to calculate the edge weights which measure the strength of the contextual relationships. And the nodes of UMGAT aggregate the intra-modal neighborhood information to further update the features of themselves. Utilizing the global information capturing ability of self-attention promotes the intra-modal long-range contextual information to be extracted. CMGAT constructs a cross-modal graph where the edges only exist between the nodes of different modalities to represent the cross-modal relationships. Similarly, the edges in CMGAT are assigned weights which are calculated via the cross-modal attention mechanism to estimate the importance of the inter-modal interaction relationships. As the graph of CMGAT is heterogeneous, the nodes involve two different modalities, so that they can provide the complementary information of the other modality for each other via the propagation and aggregation. In this process, the different modality information can fully interact with each other on the feature level, which achieves explicitly inter-modal relation modeling. To ensure the consistency of features for multimodal fusion, we define a cross-modal loss to further constrain the interaction effects of CMGATs. Moreover, an adaptive residual module is used to address the common over-smoothing in graph learning. To summarize, the main contributions of our work are

as follows:

1. A novel bi-stream graph learning framework is proposed, including the unimodal stream graph learning and the cross-modal stream graph learning. The former stream designs UMGATs to capture the intra-modal long-range contextual information, while the latter one utilizes CMGATs to learn the inter-modal interaction relationships.
2. The homogeneous and heterogeneous information learning are integrated into a uniform framework, which are all learned by GNNs. UMGATs construct homogeneous graphs with the fully connected structure to capture the long-range contextual relationships within a modality. CMGATs construct heterogeneous cross-modal graphs involving two modalities to extract the inter-modal interaction relationships as the complementary information. The two kinds of information are separately learned, which can alleviate the conflicts between different modalities and mitigate the heterogeneity issue as well.
3. A cross-modal loss is also defined to constrain the learning of CMGATs, so that it promotes the feature fusion of cross-modal. And we design an adaptive residual module to mitigate the over-smoothing common in GNNs.
4. We conduct the experiments on the two public datasets (i.e., IEMOCAP and MELD). The results show that the proposed BiGMF can outperform the state-of-the-art models.

The rest of the article is organized as follows. Section II reviews the related works of ERC from the respective of unimodal based methods and multimodal based methods. Section III provides a preliminary knowledge of the task of ERC and the proposed model BiGMF. Section IV introduces the BiGMF in details. Section V and VI present the experimental setup and experimental results on two benchmark datasets. Section VII summarizes the paper.

2. Related work

The aim of ERC is to recognize the emotional state of a speaker during a conversation. As emotions of both speakers may influence with each other, the contextual information is crucial when determining the current emotional state of the speaker. Previous recognition methods take the single modality data as input to learn the intra-modal contextual information. HiGRU [5] was a typical unimodal learning approach with bilevel GRUs. The lower level GRU modeled the word sequence of each utterance to produce the individual utterance embeddings and the upper level captured the sequential and contextual relationships between different utterances. Shen et al.[21] also proposed DialogXL on the single modality of the conversational utterances for ERC, which constructed a memory saving utterance recurrence to capture the contextual information and designed a comprehensive self-attention mechanism to learn useful intra- and inter-speaker dependencies. DialogueCRN [7] focused on extracting the emotion clues which could be regarded as the contextual information. It designed multi-turn reasoning modules to extract and integrate the emotion clues for emotion recognition. COSMIC [6] was also an utterance-level emotion recognition approach which integrated the commonsense knowledge to learn the interactions between the speakers in the conversation. Especially, the commonsense knowledge were extracted via the knowledge graph. Due to the outstanding ability of representing the complex relationships, graphs are intuitively introduced into deep learning to address various graph-related tasks, namely Graph Neural Networks (GNNs). ERC is one of the typical applications of GNNs. The conversations can be modeled via graphs that are fed to GNNs. For instance, the nodes of graph represent individual utterances, and the edges between nodes represent the dependency between the speakers of utterances. DialogueGCN [8] constructed a graph for each dialogue and defined a context window to capture the limited size of contexts, so that the graph computation could be under controllable complexity. And it used GCN with two layers of graph convolution, in which the contextual information among distant utterances could also be propagated. DAG-ERC [22] constructed a directed acyclic graph to model the conversation contexts which took the speaker identity and positional relations as the constraints. The introduction of speaker information alleviated the influence of emotional shift in conversation, but it could not avoid the shift issue in the unimodality conversation. The emotion recognition needs multi-view cues to accurately infer emotion. Fusing the information from other sources can supplement the deficiency of the unimodality emotion recognition based on conversations.

Multimodal fusion for ERC is not a brand-new research topic. Some works have shown its advantages compared with the unimodal systems. Poria et. al [12] developed a LSTM-based model to extract the intra-modal contextual features and then fuse the multimodal features, proving that the multimodal information can significantly improve the recognition accuracy. MFN [23] modeled the view-specific interactions and cross-view interactions in a multi-view conversation sequence, in which a view corresponded to a modality. CMN [24] were designed to model the emotional dynamics by analyzing the historical utterances of both speakers. The history information hidden in historical utterances was collected from the multimodal features of all utterances in a video. While ICON [25] used all the historical utterances of both speakers within a context window to learn the self-emotional influences and the inter-speaker influences via local GRUs and a global GRU, respectively. DialogueRNN [13] firstly extracted the multimodal features which were inputted to three GRUs to separately encode the speaker information, the contextual information and the historical emotion of the preceding utterances. In fusion paradigm, the aforementioned approaches equally used the concatenation to achieve the fusion of textual, audio and visual features, but the features of different modalities cannot sufficiently interact with each other.

In order to explore the ability of GNNs to process the multimodal information, some studies concatenated multimodal features as the feature representation of the nodes in the graph, including ConGCN [14] and COGMEN [15]. The performance of the approaches outperformed the unimodal ones, where the

improvement was caused by the multimodal feature concatenation. However, they neglected the exploring of the interaction between different modalities. Except for concatenation, tensor is another feasible way to represent and fuse multimodal features. A representative work is tensor fusion network (TFN) [26] which has used tensor fusion to model inter-modality dynamics in the end-to-end manner. The results of TFN also showed that the interactions between modalities should be fully exploited in the fusion. As is well known, graph is a good choice to represent complex relationships between data objects for its advantage of the structure, so that it is introduced into deep neural networks to learn the complex relations among data. Afterwards, many researchers try to model the inter-modal interaction via GNNs. MMGCN [16] constructed a completed connected graph based on all three modalities, which explored the long-range contextual relationship through the edges within the same modality nodes and the inter-modal relationship through the edges between different modalities nodes. However, MMGCN integrated the nodes of different modalities into one GNN, which neglected the heterogeneous gap in the multimodal fusion. Aiming at the heterogeneity issue, GraphMFT [18] defined three graphs, each of which only considered the information of two modalities. The improved GAT was used to capture both contextual and cross-modal information in each graph and simultaneously fuse the multimodal features. Through reducing the heterogeneous nodes within one graph, GraphMFT alleviated the heterogeneous gap to some extent. GraphCFC [19] used the same graphs as GraphMFT. Besides, it incorporated multiple subspace extractors to learn the consistency and diversity of multimodal features and get different feature representations. As to the fusion, it proposed a GAT-MLP layer to gradually fuse features of multimodality. GA2MIF [20] was a two-stage multimodal fusion approach. The first stage constructed the homogeneous graph for each modality and extracted both intra-modal local and long-range contextual information via the graph. The second stage used cross-modal attention networks to model interactions between different modalities, so that it could extract the complementary information from different modalities. Although GA2MIF was able to capture the interaction information between different modalities, the approach is less explicit than heterogeneous graph-based models and showed worse interpretability of cross-modal relationship.

Among the relation learning models, GNNs have many benefits on modeling multimodal relationships, which are the preferred choice of various deep learning methods. In the paper, we focus on utilizing the relationship modeling ability of GNNs to explicitly capture the intra-modal long-range contextual information and the inter-modal interaction relationships. Therefore, we design a bi-stream graph learning including the unimodal stream graph learning and the cross-modal stream graph learning to separately learn the two kind of information.

3. Preliminaries

We provide the preliminaries of the proposed method for ERC in the section. Firstly, a general definition for ERC is described, followed by the feature encoding used to get the raw features of three modalities. And in the end, we briefly introduce the self-attention and co-attention mechanism used in the method.

3.1. Problem definition

In ERC, a conversation can be represented as $\{u_1, u_2, u_3, \ldots, u_N\}$, where u_i represents the i-th utterance in the conversation, N is the length of the conversation, i.e., the number of utterances. Each conversation involves multiple speakers, and the speakers may exhibit different emotions in the conversation through their voice, facial expression, and language content. Therefore, u_i can be represented as $\left\{u_i^a, u_i^t, u_i^v\right\}$, where a, t, v represent the acoustic, textual, and visual modality, respectively, and u_i^a, u_i^t, u_i^v represent the raw features of the three modalities for the i-th utterance. Each utterance corresponds to an emotion, and the set of emotion labels shows the divergence in different datasets. IEMOCAP contains 6 emotion labels, while MELD contains 4 emotion labels. Given an utterance u_i , the multimodal ERC aims to predict the emotion label y_i based on the multimodal representation $\left\{u_i^a, u_i^t, u_i^v\right\}$.

3.2. Feature encoding

To better utilize the information of different modalities, a general method is to encode the unimodal data. Given the data of acoustic, textual and visual modalities, the feature encoding module employs the network with a norm layer and a fully connected layer to encode the acoustic and visual features, and uses the network with a norm layer plus two layers of bidirectional LSTM to encode the textual features, so that we can acquire features of three modalities, formulated as:

$$\begin{aligned} \mathbf{X}^a &= FC(\text{Norm}((\mathbf{U}^a))) \\ \mathbf{X}^t &= \overleftrightarrow{LSTM}(\text{Norm}(\mathbf{U}^t)) \\ \mathbf{X}^v &= FC(\text{Norm}(\mathbf{U}^v)) \end{aligned} \tag{1}$$

where, $\mathbf{U}^i \in \mathbb{R}^{N \times d_i}$, $i \in \{a, t, v\}$ is the data of a conversation with N utterances under a specific modality, d_i represents the feature dimension of the corresponding modality, and $\mathbf{X}^i$, $i \in \{a, t, v\}$ denotes the output of the unimodal feature encoding module.

3.3. Attention mechanism

Attention mechanism is introduced to quantify the strength of relationships. In the multimodal ERC, there are two types of relationships. One is the intra-modal contextual relationship and the other is the inter-modal interaction relationship, which can be measured by self-attention mechanism and co-attention mechanism, respectively. The self-attention mechanism can capture long-range dependencies in ERC, while the co-attention mechanism allows one modality to have dependency on the other modality.

The application of self-attention mechanism in Transformer and its variants is a breakthrough in the field of natural language

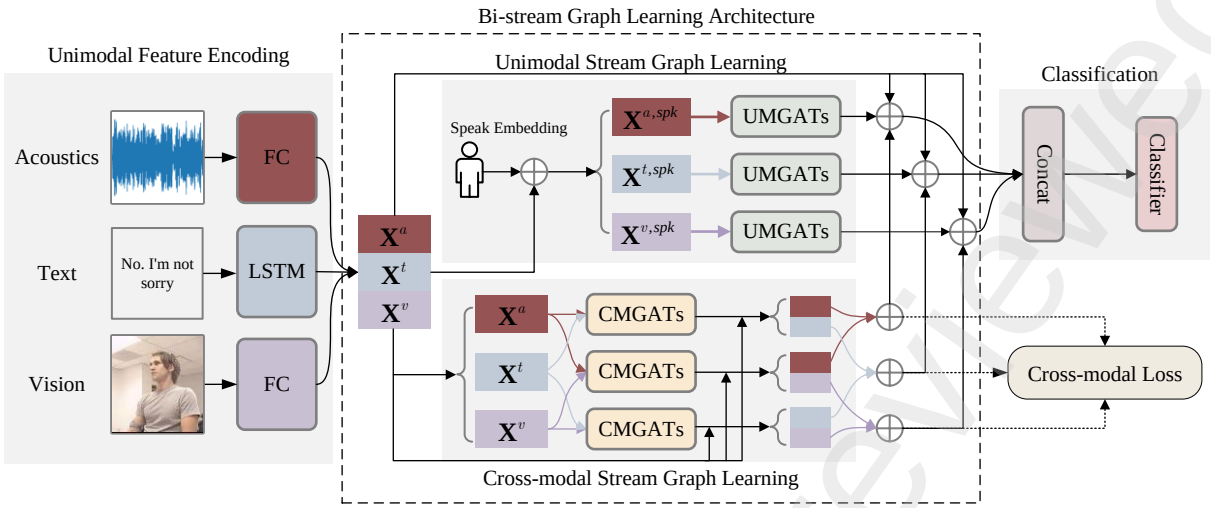


Figure 1: The framework of BiGMF.

processing (NLP) [27, 28, 29, 30, 31], which can successfully capture the global information dependency. Given input $\mathbf{X}$, two embeddings ($\mathbf{Q} = \mathbf{X}\mathbf{W}_{\mathrm{q}}$ and $\mathbf{K} = \mathbf{X}\mathbf{W}_{\mathrm{k}}$) can be generated, where $\mathbf{W}_{\mathrm{q}}, \mathbf{W}_{\mathrm{k}}$ are two projection matrices. The self-attention coefficients are defined as:

$$\alpha = \mathrm{SA}(\mathbf{Q}, \mathbf{K}) = SoftMax(\frac{\mathbf{Q} \cdot \mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}) \tag{2}$$

where d_k is the dimension of $\mathbf{Q}$ or $\mathbf{K}$. *SoftMax* is an activation function. SA is the self-attention function.

The co-attention mechanism can learn the cross-modal interactions, which has been verified by VilBERT [32] and InterBERT[33]. Given two modality inputs $\mathbf{X}^{\mathrm{A}}$ and $\mathbf{X}^{\mathrm{B}}$, four embeddings that are $\mathbf{Q}^{\mathrm{A}} = \mathbf{X}^{\mathrm{A}}\mathbf{W}_{\mathrm{q}}^{\mathrm{A}}$, $\mathbf{K}^{\mathrm{A}} = \mathbf{X}^{\mathrm{A}}\mathbf{W}_{\mathrm{k}}^{\mathrm{A}}$, $\mathbf{Q}^{\mathrm{B}} = \mathbf{X}^{\mathrm{B}}\mathbf{W}_{\mathrm{q}}^{\mathrm{B}}$, and $\mathbf{K}^{\mathrm{B}} = \mathbf{X}^{\mathrm{B}}\mathbf{W}_{\mathrm{k}}^{\mathrm{B}}$ are learned, in which $\mathbf{W}_{\mathrm{q}}^{\mathrm{A}}$,$\mathbf{W}_{\mathrm{k}}^{\mathrm{A}}$,$\mathbf{W}_{\mathrm{q}}^{\mathrm{B}}$,$\mathbf{W}_{\mathrm{k}}^{\mathrm{B}}$ are four projection matrices. Afterwards, the co-attention coefficients can be calculated as:

$$\begin{aligned} \alpha^{\mathrm{A|B}} &= \mathrm{CA}(\mathbf{Q}^{\mathrm{A}}, \mathbf{K}^{\mathrm{B}}) = SoftMax(\frac{\mathbf{Q}^{\mathrm{A}} \cdot (\mathbf{K}^{\mathrm{B}})^{\mathrm{T}}}{\sqrt{d_k}}) \\ \alpha^{\mathrm{B|A}} &= \mathrm{CA}(\mathbf{Q}^{\mathrm{B}}, \mathbf{K}^{\mathrm{A}}) = SoftMax(\frac{\mathbf{Q}^{\mathrm{B}} \cdot (\mathbf{K}^{\mathrm{A}})^{\mathrm{T}}}{\sqrt{d_k}}) \end{aligned} \tag{3}$$

where CA is the co-attention function. $\alpha^{\mathrm{A|B}}$ and $\alpha^{\mathrm{B|A}}$ measure the strength of the cross-modal interaction relationship, in which the former measures the influence of modality B to modality A and the latter measures the influence of modality A to modality B.

4. Proposed Method

Fig. 1 shows the overall framework of the proposed method, which contains a unimodal feature encoding module, a bi-stream graph learning module and the final classification for emotion recognition. The bi-stream graph learning is the core of the whole framework, constituted by a unimodal stream graph learning and a cross-modal stream graph learning which aim at learning the intra-modal long-range contextual information and the inter-modal interaction, respectively. In the section, We start with the unimodal stream graph learning and the cross-modal stream graph learning, followed by the classification module, and then end with the training scheme of the proposed BiGMF.

4.1. Unimodal stream graph learning

Different modalities refer to different context ranges, so that each modality has its specific contextual relationships and unique contextual information. Therefore, we design the unimodal stream graph learning module which takes the advantage of the powerful relationship modeling capability of graphs to extract the long-range contextual relationships and information from each of the modalities. The architecture of the module is shown in Fig. 2, including a speaker embedding component, and three unimodal graph attention networks (UMGATs) to correspondingly learn the contextual information of acoustic modality, textual modality and visual modality. Subsequently, we will introduce the speaker embedding and present the UMGATs from graph construction and graph learning.

4.1.1. Speaker embedding

Diverse speakers in the same conversation always exhibit distinct divergence in emotions. And the expressing ways may vary with different speakers. Thus, the unique information about the speaker is necessary, which can facilitate the model to learn the emotional shift of the speaker in a conversation. In the paper, we leverage the speaker information to guide the unimodal stream graph learning process, thereby enhancing the contextual feature learning for the emotional dynamics of speakers.

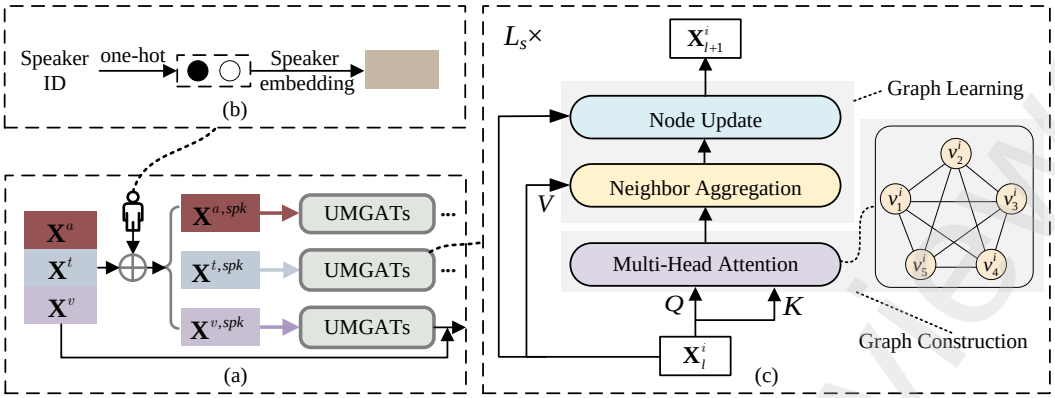


Figure 2: The illustration of Unimodal Stream Graph Learning.

As shown in Fig. 2(b), the speaker's ID is encoded to the one-hot vector, and then added it to the unimodal features, which can be represented as:

$$\begin{aligned} \mathbf{S} &= E\text{mbedding}(Spk, s) \\ \mathbf{X}_s^i &= \mathbf{X}^i + \eta^i \mathbf{S} \end{aligned} \tag{4}$$

where *Spk* is the set of speakers, and s is the number of speakers, so that $\mathbf{S}$ represents the speaker embedding. $\eta^i, i \in \{a,\ t,\ v\}$ denotes the balancing coefficient. $\mathbf{X}^i$ is the result of unimodal feature encoding, and $\mathbf{X}_s^i$ represents the modality features combined with the speaker embedding.

4.1.2. Unimodal graph construction

Graphs are constructed for each modality of one conversation, so we can get three unimodal graphs, denoted as $\mathcal{G}^i = (\mathcal{V}^i, \mathcal{E}^i, \mathcal{W}^i)$, $i \in \{a,\ t,\ v\}$, where $\mathcal{V}^i$ represents the set of nodes, $\mathcal{E}^i$ represents the set of edges, and $\mathcal{W}^i$ represents the set of edge weights. Each node of the graph denotes an utterance under a specific modality and the edge which connects two nodes represents the contextual relationship. The graph for unimodal learning is fully connected, in which any two nodes are connected by an edge. For the sake of convenience, we take the $\mathcal{G}^a$ graph as an example to introduce its construction process specifically.

Nodes: Each utterance has three modalities. The features of the corresponding acoustic modality are taken as the content of the node in graph $\mathcal{G}^a$. For instance, given a conversation containing n utterances, the $\mathcal{G}^a$ contains N nodes, i.e., $|\mathcal{V}^a| = N$.

Edges and edge weights: Edge is the connection between two nodes. And the connection actually represents the contextual relationship. Through connecting one node to all other nodes, the context can cover the whole conversation. The edges in $\mathcal{G}^a$ can represent all the contextual relationships carrying within the acoustic modality. However, the contextual relationships between utterances show different strengths. Hence, the edges are graded from strong to week according to the weights that are quantified by the self-attention mechanism. Such the self-attention weights can sufficiently represent the global information. And thus, for ERC, the weights give the importance evaluation of the contextual relationships comprehensively. Specifically, the weights of the edges are quantized as:

$$\alpha^a = \text{SA}(\mathbf{X}_s^a \mathbf{W}_q^a,\ \mathbf{X}_s^a \mathbf{W}_k^a) \tag{5}$$

where $\mathbf{X}_s^a$ is the matrix of node features in $\mathcal{G}^a$. $\mathbf{W}_q^a$ and $\mathbf{W}_k^a$ are the learnable weight matrices, which are used to project $\mathbf{X}_s^a$ into different feature representation subspaces to get query matrix $\mathbf{Q}$ and key matrix $\mathbf{K}$.

4.1.3. Graph learning for contextual information

Unimodal graph learning aims to learn contextual information for each modality. In order to capture much longer contextual information, graph learning is used, in which the contextual information can be propagated through edges. In the unimodal graph, the node features represent different segments of context in a conversation, so that we aggregate the information of all nodes to the central node through the edges for capturing the long-range contextual information. However, the propagation experiences multiple nodes that are not equally important to the contextual information, so that selecting useful nodes is crucial for inferring the state of utterances. A multi-layer graph attention learning strategy is subsequently introduced to guide neighbor aggregation.

As shown in Fig. 2(c), the graph learning in UMGATs includes neighbor aggregation and node update, in which the former summarizes the information from the context and the latter receives the contextual information to update nodes' contents. In the following, we describe the unimodal learning specifically using one layer example.

Neighbor aggregation collects features of other nodes connected to the central node, which can cover much longer range of the conversation. For central node $v_j^i, i \in \{a, t, v\}$, its neighbor

aggregation $Neibor(x^i_{s,j})$ can be defined as:

$$Neibor(x^i_{s,j}) = \Big\|_{h=1}^{H} \sum_{k \in \mathcal{V}^i_j} \alpha^{i,h}_{jk} \mathbf{W}^{i,h}_{\mathrm{v}} x^i_{s,k} \tag{6}$$

where $\alpha^{i,h}_{jk}$ is the attention weight of the h-th attention head. $\mathbf{W}^{i,h}_{\mathrm{v}}$ is the learnable weight matrix. And $x^i_{s,k}$ is the kth row of $\mathbf{X}^{\mu}_{s}$, namely the features of node v^i_k. $\mathcal{V}^i_j$ is the neighbor set of node v^i_j. $\|$ represents the concatenation operation. H denotes the number of heads.

After obtaining neighbor aggregation $Neibor(x^i_{s,j})$, node v^i_j combines the aggregation with its own node features to update its feature representation, which is formulated as:

$$x^i_{s,j}{}' = \mathbf{W}^i_2(\sigma(\mathbf{W}^i_1[(x^i_{s,j} + Neibor(x^i_{s,j})) \,\|\, (x^i_{s,j} \odot Neibor(x^i_{s,j}))])) \tag{7}$$

where $x^i_{s,j}{}'$ is the updated feature representation of node v^i_j. And $x^i_{s,j}$ represents that before update. $\mathbf{W}^i_1$ and $\mathbf{W}^i_2$ are the learnable matrices, $i \in \{a, t, v\}$. $\odot$ represents the element-wise product operation. σ is the activation function, and we use ReLU here.

Over-smoothing inevitably happens in graph learning when using the fully connected graph [34]. The direct solution is ResGCN [35]. A common way is to simply add the input features of the previous layer to those of the following layer, but unexpectedly produces lots of redundant information. Thus, we employ an adaptive residual module to dynamically preserve the initial information to address the redundancy issue mentioned above. To be specific, $\mathbf{X}^i$ is the initial node feature matrix of the node set, and $\mathbf{X}^i_{s,L}$ is the output node feature matrix through the UMGATs with L layers graph learning, so that the output contextual information $\mathbf{X}^i_u$ can be finally learned by the unimodal stream graph learning:

$$\mathbf{X}^i_u = \mathbf{X}^i_{s,L} + \mathrm{Drop}(\mathrm{Linear}(\mathbf{X}^i)) \tag{8}$$

4.2. Cross-modal stream graph learning

Although the unimodal stream graph learning has ability to capture the contextual information of each modality, it does not consider the interaction between different modalities. In the case that the information captured via unimodal learning does not match the emotion to be identified, such interaction can provide complementary information of inter-modal, which can be used to recognize the emotion from other perspectives. Moreover, inter-modal complementary information is crucial for comprehensively understanding the context. Therefore, cross-modal graphs are constructed to explicitly represent the interaction via edges, and a cross-modal graph learning is proposed to learn the complementary information from different modalities. To be specific, we design the cross-modal graph attention networks (CMGATs) to allow one modality to learn the complementary information from another modality. As shown in Fig. 3, CMGATs consists of multilayer graph construction and graph learning. In the following, we take one layer as an example to make the presentation.

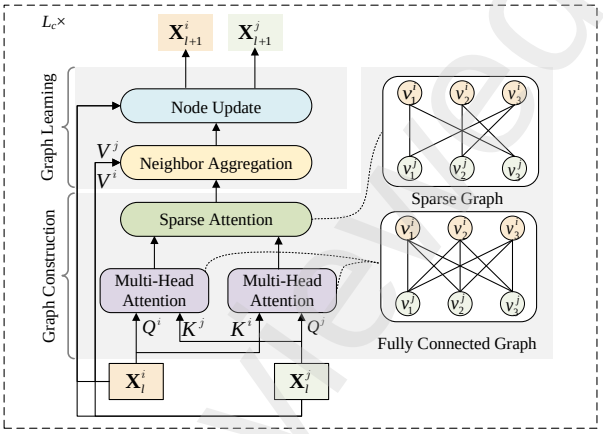


Figure 3: The illustration of CMGATs.

4.2.1. Cross-modal graph construction

Constructing the cross-modal graph is the first step of CMGATs. Unlike the unimodal graph, the cross-modal graphs have changes in both nodes and edges. Each cross-modal graph involves two modalities. Each of the nodes represents the utterance under a certain modality. While the edge only connects the two nodes of different modalities and denotes the interaction between the two different modalities. In this way, three cross-modal graphs are constructed: the $a-t$ graph, $a-v$ graph, and $t-v$ graph, denoted as $\mathcal{G}^{a\cup t} = (\mathcal{V}^{a\cup t}, \mathcal{E}^{a\cup t}, \mathcal{W}^{a|t}, \mathcal{W}^{t|a})$, $\mathcal{G}^{a\cup v} = (\mathcal{V}^{a\cup v}, \mathcal{E}^{a\cup v}, \mathcal{W}^{a|v}, \mathcal{W}^{v|a})$ and $\mathcal{G}^{t\cup v} = (\mathcal{V}^{t\cup v}, \mathcal{E}^{t\cup v}, \mathcal{W}^{t|v}, \mathcal{W}^{v|t})$. The following will take $a-t$ graph as the example to explain the specific construction process.

Nodes: In cross-modal graph $\mathcal{G}^{a\cup t}$, each utterance is treated as two nodes which represent acoustic modality and textual modality, respectively. Thus, for a dialogue with N utterances, $\mathcal{G}^{a\cup t}$ graph contains $2N$ nodes and the feature of each node is represented by the corresponding modality. It can be observed that $\mathcal{V}^{a\cup t} = \mathcal{V}^a \bigcup \mathcal{V}^t$.

Edges and Edge Weights: Edges in cross-modal graphs only connect different modality nodes. For instance, we connect an acoustic node to all textual modality nodes, which means all textual nodes have interactive relationships with the acoustic node. The interactions are assumed to be bidirectional, and have discrepancy in relation strength. Therefore, the co-attention mechanism is introduced to measure such bidirectional interaction by calculating the edge weights. The weights of edges in $\mathcal{G}^{a\cup t}$ graph can be computed as:

$$\begin{aligned} \alpha^{a|t} &= \mathrm{CA}(\mathbf{Q}^a, \mathbf{K}^t) \\ \alpha^{t|a} &= \mathrm{CA}(\mathbf{Q}^t, \mathbf{K}^a) \end{aligned} \tag{9}$$

where $\mathbf{Q}^i$ and $\mathbf{K}^i$ represent query matrix and key matrix of modality i, $i \in \{a, t, v\}$. $\alpha^{a|t}$ measures the relation strength of textual modality nodes to the acoustic modality nodes, and $\alpha^{t|a}$ measures the that of acoustic modality nodes to the textual ones.

The node of one modality is fully connected to other modality nodes, so that the unnecessary connections inevitably exist in the cross-modal graph, which hinders the learning of effective complementary information. Therefore, we pick out the more representative ones from the constructed edges by making the co-attention weight matrix sparse. Concretely, the top K weights in co-attention weight matrix are reserved, and the remaining values are set to zero. The process is dynamically adjusted with node updating in cross-modal graph learning. In this way, the edges with weight zero are removed, and a sparse cross-modal graph is constructed in which each node has K neighbors.

4.2.2. Graph learning for complementary information

The cross-modal graphs focus on learning the complementary information from multimodal data for emotion recognition. Through edges across different modalities, features of one modality can be propagated to another modality. As shown in Fig. 3, the cross-modal graph learning consists of neighbor aggregation and node update. The former is to gather the features from other modality, and the latter operation is to fuse the features of two modalities.

Taking the $a-t$ graph as the example, we firstly define K neighbors of central node $v_i^{a\cup t} \in \mathcal{V}^{a\cup t}$ as $N_K\left(v_i^{a\cup t}\right)$. Then, the aggregated neighbor features $Neibor(x_i^{a\cup t})$ of K neighbors can be learned by:

$$Neibor(x_i^{a\cup t}) = \begin{cases} \Big\Vert_{h=1}^{H} \sum\limits_{N_K(v_i^{a\cup t})} \alpha_{ij}^{a|t,h} \mathbf{W}_{\mathrm{v}}^{t|a,h} x_j^t, & v_i^{a\cup t} \in \mathcal{V}^a \\ \Big\Vert_{h=1}^{H} \sum\limits_{N_K(v_i^{a\cup t})} \alpha_{ij}^{t|a,h} \mathbf{W}_{\mathrm{v}}^{a|t,h} x_j^a, & v_i^{a\cup t} \in \mathcal{V}^t \end{cases} \tag{10}$$

where $\alpha_{ij}^{a|t}$ and $\alpha_{ij}^{t|a}$ denote the co-attention coefficients. x_j^t and x_j^a are the j-th row of $\mathbf{X}^t$ and $\mathbf{X}^a$, representing the features of the textual modality node and acoustic modality node for utterance u_j. $\mathbf{W}_{\mathrm{v}}^{t|a,h}$ and $\mathbf{W}_{\mathrm{v}}^{a|t,h}$ are learnable matrices. $\mathcal{V}^a$ is the set of all acoustic nodes and $\mathcal{V}^t$ is the set of all textual nodes. $\|$ represents the concatenation operation. H is the number of attention heads.

It should be noted that all the neighbors of the central node belong to the same modality, so there is no heterogeneous fusion issue when aggregating the neighbor features. In this way, the features of the neighbors can provide the complementary information for the central node from different modality. Moreover, such cross-modal learning paradigm successfully avoids the heterogeneous issue caused by multimodal fusion within one graph in many approaches.

After that, new features $x_i^{a\cup t\prime}$ of node $v_i^{a\cup t}$ can be update by:

$$x_i^{a\cup t\prime} = \mathbf{W}_2^{a\cup t}(\sigma(\mathbf{W}_1^{a\cup t}[(x_i^{a\cup t} + Neibor(x_i^{a\cup t})) \,\|\, (x_i^{a\cup t} \odot Neibor(x_i^{a\cup t}))])) \tag{11}$$

where $x_i^{a\cup t}$ denotes the features of node $v_i^{a\cup t}$ before the update. σ is the activation function and here we use ReLU. $\mathbf{W}_1^{a\cup t}$ and $\mathbf{W}_2^{a\cup t}$ are learnable matrices. Therefore, $\mathbf{X}_L^{a\cup t}$ contains the complementary information of textual modality to acoustic modality as well as that of acoustic modality to textual modality.

Feature matrix $\mathbf{X}_L^{a\cup t}$ learned from $\mathcal{G}^{a\cup t}$ can be divided into $\mathbf{X}_L^{a|t}$ and $\mathbf{X}_L^{t|a}$ in terms of the modality to which it belongs. The same operation is conducted on $\mathcal{G}^{a\cup v}$ and $\mathcal{G}^{t\cup v}$ to get feature matrices $\mathbf{X}_L^{a|v}$ and $\mathbf{X}_L^{v|a}$, $\mathbf{X}_L^{t|v}$ and $\mathbf{X}_L^{v|t}$.

Till now, the graph nodes can receive the complementary information of one additional modality. However, the remaining modality need to be further fused into the current features. After completing the inter-modal interaction between two modalities, the features involving the same modality are fused subsequently to obtain more comprehensively complementary information. Through the inter-modal interaction via the cross-modal stream graph learning, the feature matrix of one modality contains the information from other two modalities.

To preserve modality-specific features during the cross-modal graph learning, an adaptive residual module is applied to add the initial features of each modality to the output of CM-GATs. Then, the ultimated output cross-modal features, each of which contain modality-specific information and complementary information from the other two modalities, can be represented as:

$$\begin{aligned} \mathbf{X}_c^a &= \mathbf{X}_L^{a|t} + \mathbf{X}_L^{a|v} + \mathrm{Drop}(\mathrm{Linear}(\mathbf{X}^a)) \\ \mathbf{X}_c^t &= \mathbf{X}_L^{t|a} + \mathbf{X}_L^{t|v} + \mathrm{Drop}(\mathrm{Linear}(\mathbf{X}^t)) \\ \mathbf{X}_c^v &= \mathbf{X}_L^{v|a} + \mathbf{X}_L^{v|t} + \mathrm{Drop}(\mathrm{Linear}(\mathbf{X}^v)) \end{aligned} \tag{12}$$

where $\mathbf{X}^i, i \in \{a,t,v\}$ is the initial input features of the corresponding modality.

Moreover, two kinds of losses are designed to guide the cross-modal stream graph learning. One is the cross-entropy loss. Specifically, since $\mathbf{X}_c^a$, $\mathbf{X}_c^t$ and $\mathbf{X}_c^v$ aggregate the features of the other two modalities, they are expected to be able to achieve a good emotion classification separately. For utterance j with features $x_{c,j}^i$, $i \in \{a,t,v\}$, its emotion probability can be calculated by a fully connected layer fc:

$$p_j^i = SoftMax(fc(x_{c,j}^i)) \tag{13}$$

Afterwards, cross-entropy loss function $CE(\cdot)$ is employed to compute the loss between the classified result p_j^i and its ground-truth y_j.

$$\mathcal{L}_{ce}^i = CE\left(p_j^i, y_j\right) \tag{14}$$

The other is the consistency loss used to further ensure the similarity between $\mathbf{X}_c^a$, $\mathbf{X}_c^t$ and $\mathbf{X}_c^v$, which is specific as:

$$\mathcal{L}_{cl} = \left\|\mathbf{X}_c^a - \mathbf{X}_c^t\right\|_2 + \left\|\mathbf{X}_c^a - \mathbf{X}_c^v\right\|_2 + \left\|\mathbf{X}_c^t - \mathbf{X}_c^v\right\|_2 \tag{15}$$

where $\|\cdot\|_2$ represents the L2-norm.

4.3. Fusion learning of unimodal and cross-modal features

Unimodal features contain the contextual information and cross-modal features carry the complementary information. For each modality, fusing these features can improve the information in the respective view. And the multimodal features fusion can bring a comprehensive information from multiple views. So, the features learned by the unimodal stream graph learning are added to those learned by the cross-modal stream graph

learning which are related to the current modality. Then, the final fusion features can be achieved by the concatenation. The operation can be described as:

$$\mathbf{X} = (\mathbf{X}_u^a + \mathbf{X}_c^a) \parallel (\mathbf{X}_u^t + \mathbf{X}_c^t) \parallel (\mathbf{X}_u^v + \mathbf{X}_c^v) \tag{16}$$

For each utterance u_i, the fused features $x_i \in \mathbf{X}$ are inputted into a fully connected layer *fc* to predict the emotional label:

$$p_i = SoftMax(fc(x_i)) \tag{17}$$

where p_i denotes the predicted probability vector of utterance u_i. Then, the loss between the predicted result and its ground-truth y_i is computed by the cross-entropy function $CE(\cdot)$ specified as:

$$\mathcal{L}^{avt} = CE(p_i, y_i) \tag{18}$$

Conclusively, the overall training objective function is the combination of the above loss functions, which can be formalized as:

$$\mathcal{L} = \mathcal{L}^{avt} + \zeta^a \mathcal{L}_{ce}^a + \zeta^t \mathcal{L}_{ce}^t + \zeta^v \mathcal{L}_{ce}^v + \gamma \mathcal{L}_{cl} \tag{19}$$

where ζ^a, ζ^t, ζ^v and γ are the trade-off weights to balance the fusing process of different modalities.

4.4. Training of BiGMF

The specific training of BiGMF is presented in Algorithm 1. Given three modalities, the first step of BiGMF is to preprocess multimodal data, as indicated in 1st line. Then, the raw features are further learned by the unimodal network and cross-modal network to respectively capture the intra-modal long-range contextual information of unimodal and get the complementary information of interaction among different modalities, which corresponds to line 2 to line 13 and line 14 to line 26. Furthermore, the unimodal features and the cross-modal features are fused as the ultimated representation which will be used for classification. At last, the overall training loss is calculated to update the proposed BiGMF.

Algorithm 1 BiGMF

Input: unimodal raw features $U^i,\ i \in \{a,\ t,\ v\}$.
Output: The final fused multimodal features.

1: Encode the unimodal raw features by (1) and get $\mathbf{X}^i,\ i \in \{a,\ t,\ v\}$
2: Add the speaker information by (4) and get $\mathbf{X}_s^i,\ i \in \{a,\ t,\ v\}$
3: **For** $i \in \{a, t, v\}$ **do**
4: Construct unimodal graph $\mathcal{G}^i\left(\mathcal{V}^i, \mathcal{E}^i, \mathcal{W}^i\right)$
5: **For** $l = 0 \ldots L_s$ **do**
6: **For** $v \in \mathcal{V}^i$ **do**
7: Aggregate the features of neighboring nodes of node via Equation (6)
8: Update the features of node via Equation (7)
9: **End**
10: **End**
11: get the final node feature matrix $\mathbf{X}_{s,L}^i$
12: Apply an adaptive residual module to $\mathbf{X}_{s,L}^i$ and get $\mathbf{X}_u^i, i \in \{a, t, v\}$ via Equation (7)
13: **End**
14: **For** $(i, j) \in \{(a, t), (a, v), (t, v)\}$ **do**
15: Construct cross-modal graph $\mathcal{G}^{i\cup j}\left(\mathcal{V}^{i\cup j}, \mathcal{E}^{i\cup j}, \mathcal{W}^{i|j}, \mathcal{W}^{j|i}\right)$
16: **For** $l = 0 \ldots L_c$ **do**
17: **For** $v \in \mathcal{V}^{i\cup j}$ **do**
18: Aggregate the features of neighboring nodes of node via Equation (10)
19: Update the features of node via Equation (11)
20: **End**
21: **End**
22: get the final node feature matrix $\mathbf{X}_L^{i\cup j}$
23: Divide the $\mathbf{X}_L^{i\cup j}$ and get $\mathbf{X}_L^{i|j}$ $\mathbf{X}_L^{j|i}$
24: Apply an adaptive residual module to $\mathbf{X}_L^{a|t}$, $\mathbf{X}_L^{t|a}$ and get $\mathbf{X}_c^{i|j}$, $\mathbf{X}_c^{j|i}$
25: **End**
26: get three cross-modal output features $\mathbf{X}_c^a$, $\mathbf{X}_c^t$ and $\mathbf{X}_c^v$ via Equation (12)
27: Calculate cross-modal losses $\mathcal{L}_{ce}^i,\ i \in \{a, t, v\}$ and $\mathcal{L}_{cl}$ by (14) and (15)
28: Fuse the context features and interaction features via (16)
29: Calculate final classification loss by (18)
30: Combine multiple losses by (19)
31: Back propagation and update parameters in BiGMF

5. Experiments

5.1. Dataset

We verify the proposed model using two benchmark ERC datasets, IEMOCAP [36] and MELD [37], which are multimodal datasets including acoustic, textual, and visual modalities. The raw feature extraction methods for the datasets are consistent with those in MMGCN, which use the OpenSmile toolkit [38], TextCNN [39], and DenseNet [40] to extract acoustic, textual, and visual features, respectively. Table 1 sums up the characteristics of the two datasets used in the paper.

The IEMOCAP dataset records dyadic dialogues of 10 actors who are asked to perform selected emotional scripts and improvised hypothetical scenarios to elicit specific types of emotions including happy, sad, neutral, angry, excited and frustrated, which contains approximately 12 hours of data, with a total of 151 dialogues and 7433 utterances.

The MELD dataset extends the Emotion Lines dataset to the multimodal scenario. It contains 13708 utterances from 1433 dialogues from TV series *Friends*. Each dialogue includes three or more speakers and each utterance is annotated with one of seven emotion labels: anger, disgust, fear, joy, neutral, sadness and surprise. Due to the larger number of speakers in the dataset, the classification is more challenging compared to dyadic dialogues. At the same time, the class imbalance brings more difficulty into MELD for ERC task.

Table 1: The summary of IEMOCAP and MELD

Dataset	Dialogues			Utterances			Classes	Dimensions		
	train	valid	test	train	valid	test		acoustic	textual	visual
IEMOCAP	120		31	5810		1623	6	1582	100	342
MELD	1039	114	280	9989	1109	2610	7	300	600	342

5.2. Implementation details

We used Adam optimizer to train the model on both IEMOCAP and MELD datasets. Table 2 shows the hyperparameter settings in the experiment. The dropout rate is set to 0.5 and the neighbor size K in the cross-modal graph is set to 5. For IEMOCAP dataset, we use 1 layer for UMGATs (L_s) and 4 layer for CMGATs (L_c), respectively. The learning rate is set to 0.00003. The regularization factor is 0.00001. The batch size is 16. And the speaker coefficients are set as 0.4, 0.04 and 0.004, respectively. For MELD dataset, both UMGATs (L_s) and CMGATs(L_c) have 1 network layer. The learning rate is 0.0003. The regularization factor is zero. The batch size is 32. And the speaker coefficients are set as 0.8, 0.5 and 0.5, respectively. As to the class-imbalance in MELD, we follow MMGCN to use the focal loss when training BiGMF on MELD.

5.3. Baselines

To verify the proposed BiGMF, we compare with several baseline models including non-graph methods and graph-based methods. The non-graph methods include BC-LSTM [12], TFN [26]and DialogueRNN [13], while the graph-based methods include DialogueGCN [8], MMGCN [16], GraphMFT [18], GraphCFC [19] and GA2MIF [20].

BC-LSTM uses LSTM networks to extract contextual features and fuse multimodal features, but fails to consider speaker information, which is important for modeling contextual dependencies. TFN learns both the intra-modal and inter-modal dynamics end-to-end without capturing the context from surrounding utterances. DialogueRNN takes the characteristic of the speaker into account and use this information for capturing the finer context. DialogueGCN uses the graph convolution network where the nodes represent individual utterances to address the context propagation issue in RNN-based methods. MMGCN simultaneously models intra-speaker context dependency and inter-speaker dependency in the same graph where nodes represent different modalities of different utterances. GraphMFT designs three graphs and each contains two modalities to alleviate data heterogeneity influence, and uses the improved GATs to facilitate the interaction learning of current utterance with intra-modal and inter-modal utterances. GraphCFC constructs the graph in the same way as GraphMFT and designs the subspace extractors to get two different feature representations for each modality, which are pairwise input to the GNN to gradually fuse the features. GA2MIF is a two-stage model, where the first step adopts three GATs to respectively model the contextual information of three modalities and the second step introduces the cross-attention to model the cross-modal interactions avoiding the use of heterogeneous graph. GA2MIF is currently the best model for multimodal ERC.

6. Results and discussions

In this section, we report and discuss the results of our experiments on IEMOCAP and MELD datasets. At first, we compare the proposed BiGMF with all baseline models. Then, the experiments under different settings of BiGMF are given. In addition, we analyze the limitations of BiGMF.

6.1. Comparative experiments

Table 3 presents the comparison results of BiGMF and other models on two datasets, all of which are carried out in the unified multimodal setting. As shown in the table, for IEMOCAP dataset, the accuracy and weighted-average F1 score of BiGMF are both improved by about 10% compared to BC-LSTM, indicating that the graph-based method can better capture contextual information compared to traditional recurrent-based methods, which can also be confirmed from the improvement of BiGMF over DialogueRNN. Compared with TFN, models based on heterogeneous graphs have achieved significant improvements, reflecting the advantages of modeling cross-modal relationships with heterogeneous graphs. Compared with MMGCN in accuracy and weighted-average F1 score, GraphMFT shows an improvement of 1.97% and 2.39%, GraphCFC improves by 3.2% and 3.23%, and our proposed BiGMF achieves the highest improvement of 3.82% and 3.8%, proving that BiGMF handles the heterogeneity issue better. Noting that BiGMF has the same accuracy as GA2MIF, but a lower weighted-average F1 score. The reason is that GA2MIF uses cross-modal attention mechanism instead of heterogeneous graphs to model the cross-modal relationships, avoiding the direct fusion of features of different modalities and thus is not affected by the multimodal heterogeneity. However, the model interpretability of GA2MIF is influenced by its implicit exploration of relationships between modalities, while BiGMF can explicitly model the cross-modal relationships with heterogeneous graphs as well as has strong interpretability.

For MELD dataset, BiGMF outperforms other approaches in both accuracy and weighted-average F1 score, with accuracy reaching to 62.03%. However, we find that the discrepancy on performance among all the models are relatively small and the improvement made by BiGMF is not very significant. It should be noted that approximately 42% of the utterances in MELD are shorter than five words, which restricts the context representation. Meanwhile, the utterances are not strictly continuous but the timestamps are sorted in an increasing order. The two limitations above may disturb the contextual relationship modeled by graphs. Simultaneously, the inherent background noise within MELD also impacts the feature learning.

Based on the above analysis, MELD has much shorter dialogues. The fusion will bring more abundant information being

Table 2: The hyperparameter settings in experiments.

Dataset	L_s	L_c	K	Learning rate	Regularization factor	Batch size	Dropout rate	η^a, η^t, η^v
IEMOCAP	1	4	5	0.00003	0.00001	16	0.5	0.4, 0.04, 0.004
MELD	1	1	5	0.0003	0	32	0.5	0.8, 0.5, 0.5

Table 3: Comparison with the baseline models on IEMOCAP and MELD datasets.

Model	IEMOCAP								MELD	
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	wa-F1	Acc	wa-F1
	F1	F1	F1	F1	F1	F1				
BC-LSTM	32.63	70.34	51.14	63.44	67.91	61.06	59.58	59.10	59.62	56.80
TFN	38.29	64.12	51.43	55.34	57.95	57.58	56.12	55.54	59.77	57.01
DialogueRNN	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75	60.31	57.66
DialogueGCN	47.01	80.88	58.71	66.08	70.97	61.21	65.54	65.04	58.62	56.36
MMGCN	40.00	76.19	62.87	70.12	74.60	62.48	65.93	65.68	59.69	57.67
GraphMFT	45.99	83.12	63.08	70.30	76.92	63.84	67.90	68.07	61.30	58.37
GraphCFC	43.08	84.99	64.70	71.35	78.86	63.70	69.13	68.91	61.42	58.86
GA2MIF	46.15	84.50	68.38	70.29	75.99	66.49	69.75	70.00	61.65	58.94
BiGMF	48.56	82.45	66.15	70.39	77.40	65.76	69.75	69.48	62.03	58.95

beneficial to emotion recognition. While GA2MIF has longer dialogues, and understanding the emotion flow is more dependent on the contextual features. The direct fusion may introduce interference to the contextual information. BiGMF has more improvement on MELD than GA2MIF.

6.2. BiGMF under different modality setting

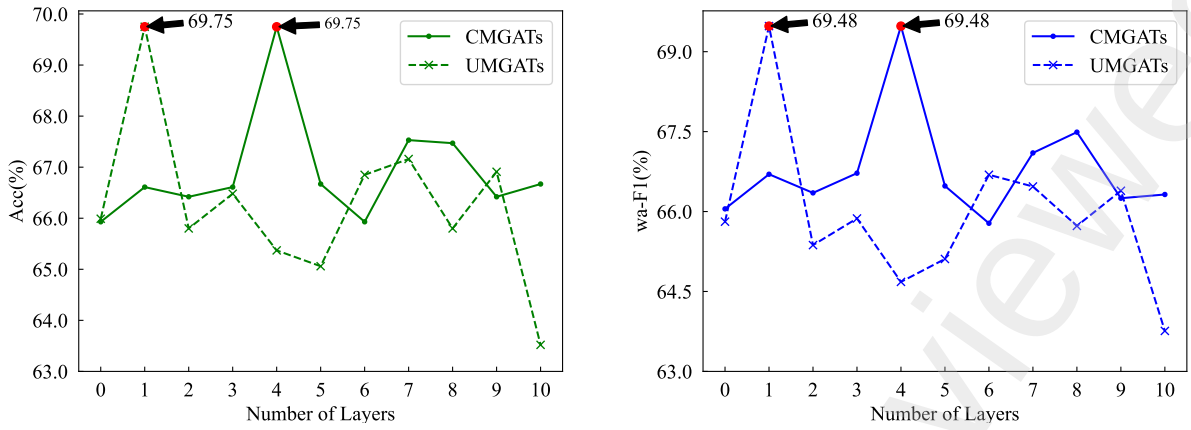
Table 4: The performance of BiGMF under different modality settings.

Modality setting	IEMOCAP		MELD	
	Acc	wa-F1	Acc	wa-F1
$a + t$	65.74	66.01	60.73	57.55
$a + v$	45.78	45.26	49.62	42.43
$t + v$	63.03	62.76	60.46	57.33
$a + t + v$	69.75	69.48	62.03	58.95

To investigate the effect of different modality settings on BiGMF, we carry out experiments under different modality combination on the two datasets. From Table 4, it can be clearly observed that the setting with three modalities obviously outperforms the settings with two-modality combination, proving that more modalities can provide more information for emotion recognition. Among the results under the different combination of two modalities, the case with the acoustic-textual combination achieves the best, while the case with the acoustic-visual combination has the worst performance. For ERC, the acoustic modality and textual modality carry abundant information about emotions which can complement each other well. While the visual modality is materialized as images in which the subtle expression is difficult to capture or easy to misunderstand, resulting into the conflict between acoustic and visual modalities and introducing more noise to emotion recognition. Moreover, under the settings fusing the textual modality, the performance becomes better. It further demonstrates the importance of the textual modality, which carries more contextual information within it.

6.3. Impact of network layers

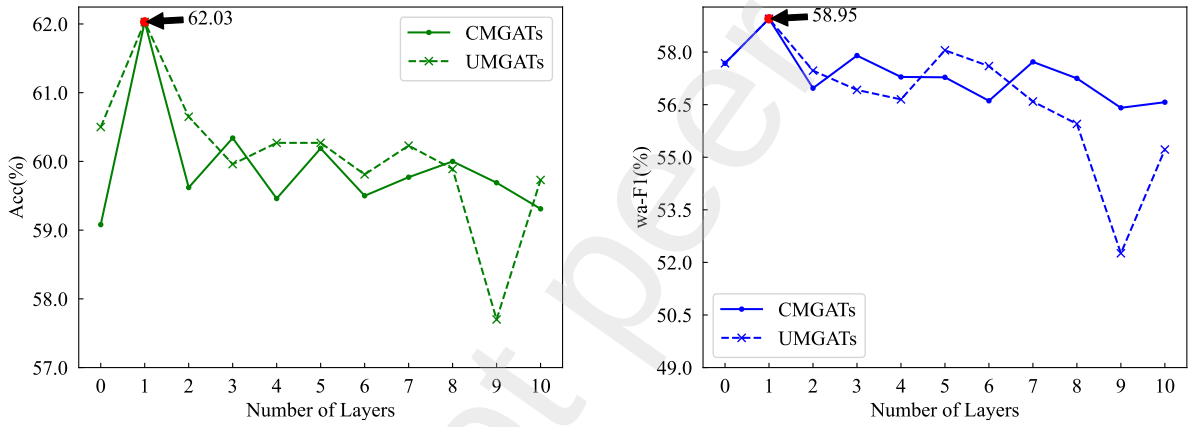
Over-smoothing issue always occurs with the deepening of graph network, while too shallow network also cannot fully learn the node feature representation. It is critical to select proper network layers for BiGMF. In the following, we study the impact of number of network layers on the proposed BiGMF. BiGMF contains two types of graph learning networks, including UMGATs and CMGATs. The former is used to learn the intra-modal long-range contextual relationships, while the latter is used to learn the inter-modal interaction relationships. Fig. 4 and 5 show the performance of BiGMF with different numbers of layers in UMGATs and CMGATs, respectively. Considering that the two graph learning networks jointly affect the performance of the model, we fix the number of layers of one and change the number of layers of the other. Fig. 4(a) shows the variation of accuracy with the layer numbers. The solid line displays the impact of the CMGATs network layers on BiGMF when the number of UMGATs layers is fixed at 1, and the dashed line displays the impact of the UMGATs network layers on BiGMF when the number of CMGATs layers is fixed at 4. Fig. 4(b) shows the variation of the weighted-average F1 score under the same setting. It can be observed that the model achieves the best performance on IEMOCAP when the number of UMGATs layers is 1 and the number of CMGATs layers is 4. Similarly, Fig. 5(a) and Fig. 5(b) show the variation of accuracy and weighted-average F1 score on MELD. When studying the impact of one graph learning network layers on BiGMF, the layer number of the other graph learning network is fixed at 1. It can be observed that the model achieves the best performance when the layer numbers of UMGATs and CMGATs both equal to1. The graph constructed in UMGATs is a fully connected



(a) The variation of accuracy score

(b) The variation of the weighted-average F1 score

Figure 4: The performance of BiGMF with different number of network layers on IEMOCAP dataset



(a) The variation of accuracy score

(b) The variation of the weighted-average F1 score

Figure 5: The performance of BiGMF with different number of network layers on MELD dataset

graph, and thus the one-layer structure is sufficient to capture the long-range contextual information. Considering the complexity of MELD dataset, multiple layers of graph networks for CMGATs will aggregate the errors caused by the multi-hop propagation during feature learning, so that the one-layer network is enough and beneficial to reduce such errors.

6.4. Ablation study

In order to verify the effect of each module, we do the ablation study on IEMOCAP and MELD datasets. Table 5-7 show the results of ablation study by gradually adding each module to the baseline model.

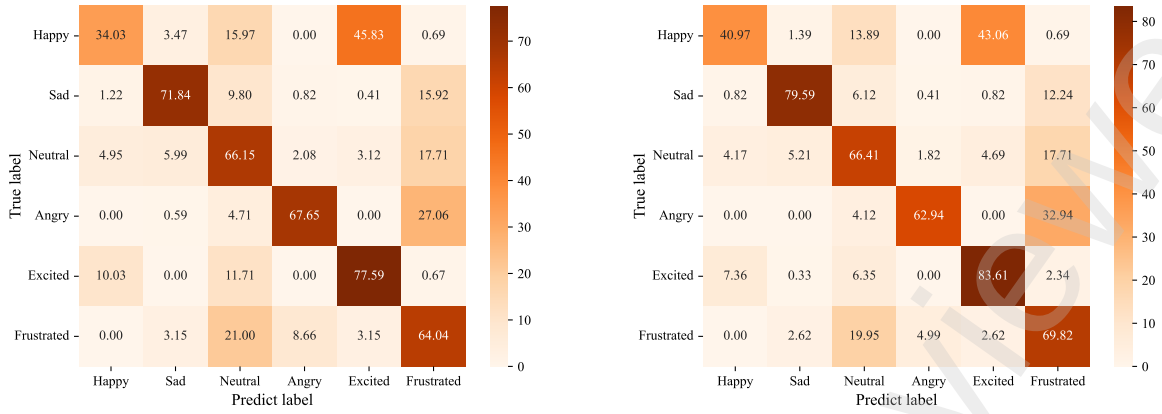
Table 5 shows the performance of the model with or without UMGATs, which is used to learn the long-range contextual information of each modality. It can be observed that the accuracy and weighted-average F1 score of BiGMF on IEMOCAP decrease by 4.07% and 3.76% without UMGATs, respectively. The decline also exists in the results of MELD, but is less than those of IEMOCAP. Thus, UMGATs is important to learn intra-modal contextual information. And IEMOCAP has more contextual information than MELD to be captured.

Table 5: Ablation study of unimodal stream graph learning.

UMGATs	IEMOCAP		MELD	
	Acc	wa-F1	Acc	wa-F1
-w/o	65.68	65.72	61.19	57.76
-w	69.75	69.48	62.03	58.95

Table 6: Ablation study of cross-modal stream graph learning.

CMGATs	Cross-modal Loss	IEMOCAP		MELD	
		Acc	wa-F1	Acc	wa-F1
-w/o	-w/o	66.05	66.11	61.11	57.61
-w	-w/o	67.34	67.05	60.54	57.30
-w	-w	69.75	69.48	62.03	58.95



(a) MMGCN (b) BiGMF

Figure 6: Comparison of confusion matrices between MMGCN and BiGMF

Table 6 shows the ablation study of cross-modal stream graph learning with or without CMGATs and cross-modal loss. To verify the effect of them, we gradually add the components to the baseline. For the baseline without CMGATs and cross-modal loss, the accuracy and weighted-average F1 score of BiGMF on IEMOCAP decrease by 3.7% and 3.37%, which manifests the necessity of the complementary information learned by the cross-modal stream graph learning. As to MELD, the accuracy and weighted-average F1 score are 60.54% and 57.30% in the case that only CMGATs module is retained. With the cross-modal loss, the CMGATs module work much better, indicating that the loss can effectively guide the learning of CMGATs.

Table 7: Ablation study of adaptive residual.

Adaptive Residual	IEMOCAP		MELD	
	Acc	wa-F1	Acc	wa-F1
-w/o	66.97	66.81	61.15	57.47
-w	69.75	69.48	62.03	58.95

Table 7 shows the ablation study of adaptive residual module. On both IEMOCAP and MELD, the performance of adaptive residual module is superior to that without adaptive residual module, verifying the necessity of the adaptive residual.

6.5. Confusion matrix of emotion recognition

At last, we give the results and analysis of the comparison between MMGCN and our proposed BiGMF based on the confusion matrix of emotion recognition.

Fig. 6 shows the confusion matrix of MMGCN and BiGMF on IEMOCAP dataset. From Fig. 6, BiGMF reduces the misclassification probability of the sad as frustrated, from 15.92% of MMGCN to 12.24%. Similarly, the probability of the emotion excited misclassified as happy is declined from 10.03% of MMGCN to 7.36%. And it is worth noting that BiGMF alleviates the non-Neutral issue caused by unimodal information mentioned in GraphMFT [18]. Specifically, MMGCN always misclassifies other ground truth emotions as neutral. In Fig.6, the misclassification probabilities of happy, sad and excited reach to 15.97%, 9.80% and 11.71%, respectively. While BiGMF efficiently depresses the misclassification probabilities, which can get the improvements of 2.08%, 3.68% and 5.36%, respectively. Hence, BiGMF can better fuse the multimodal information to improve the performance of ERC.

However, there still have limitations in current ERC models, including ours. The current ERC models identically suffers from undistinguishable issue when confronting similar emotions, such as happy and excited. As shown in Fig. 6, MMGCN and BiGMF misclassify most of the ground-truth happy as excited, even over 40%. The happy emotion with strong degree has the similar external expression with emotion excited. Even though, BiGMF still has better discrimination ability than MMGCN.

7. Conclusion

To fully fuse the information of different modalities, a novel multimodal fusion approach named Bi-stream Graph Learning based Multimodal Fusion (BiGMF) is proposed, which provides an innovative scheme to parallelly model the long-range contextual relationship of each modality and the interaction of every two modalities by the unimodal graph and the cross-modal graph, respectively. In the two types of graph, the contextual information and the complementary information are both propagated through edges. And both of the information is fused in the process of bi-stream graph learning. In addition, the adaptive residual module is designed to alleviate the over-smoothing in unimodal graph learning and preserve the modality-specific features during the cross-modal graph learning. To further guide the learning of cross-modal graph, we define the cross-modal loss to guide the training of BiGMF. The proposed BiGMF has been experimentally verified on two public datasets of ERC, and the results show that the method provides an effective way to fuse the multimodal information for the emotion recognition task in conversation.

In future work, we will focus on the issues of similar-emotion and non-neutral, and further study how to efficiently encode the speaker information on the basis of multimodal fusion to capture the emotion flow in a conversation. Also, the scheme is expected to be achieved with the help of powerful relationship modeling capability of GNNs.

Acknowledge

This work was supported by National Natural Science Foundation of China (No. 62006233, 62173063) and National Key R&D Program of China (2019YFE0118500).

References

[1] H. Zhou, M. Huang, T. Zhang, X. Zhu, B. Liu, Emotional chatting machine: Emotional conversation generation with internal and external memory, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[2] R. L. Rosa, G. M. Schwartz, W. V. Ruggiero, D. Z. Rodríguez, A knowledge-based recommendation system that includes sentiment analysis and deep learning, IEEE Transactions on Industrial Informatics 15 (2018) 2124–2135.

[3] A. Chatterjee, K. N. Narahari, M. Joshi, P. Agrawal, Semeval-2019 task 3: Emocontext contextual emotion detection in text, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 39–48.

[4] A. Bhavan, P. Chauhan, R. R. Shah, et al., Bagged support vector machines for emotion recognition from speech, Knowledge-Based Systems 184 (2019) 104886.

[5] W. Jiao, H. Yang, I. King, M. R. Lyu, Higru: Hierarchical gated recurrent units for utterance-level emotion recognition, arXiv preprint arXiv:1904.04446 (2019).

[6] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, S. Poria, Cosmic: Commonsense knowledge for emotion identification in conversations, arXiv preprint arXiv:2010.02795 (2020).

[7] D. Hu, L. Wei, X. Huai, Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations, arXiv preprint arXiv:2106.01978 (2021).

[8] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, arXiv preprint arXiv:1908.11540 (2019).

[9] F. Huang, X. Li, C. Yuan, S. Zhang, J. Zhang, S. Qiao, Attention-emotion-enhanced convolutional lstm for sentiment analysis, IEEE transactions on neural networks and learning systems 33 (2021) 4332–4345.

[10] L. Yi, M.-W. Mak, Improving speech emotion recognition with adversarial data augmentation network, IEEE transactions on neural networks and learning systems 33 (2020) 172–184.

[11] I. A. Essa, A. P. Pentland, Coding, analysis, interpretation, and recognition of facial expressions, IEEE transactions on pattern analysis and machine intelligence 19 (1997) 757–763.

[12] 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.

[13] 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, volume 33, 2019, pp. 6818–6825.

[14] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, G. Zhou, Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations., in: IJCAI, 2019, pp. 5415–5421.

[15] A. Joshi, A. Bhat, A. Jain, A. V. Singh, A. Modi, Cogmen: Contextualized gnn based multimodal emotion recognition, arXiv preprint arXiv:2205.02455 (2022).

[16] J. Hu, Y. Liu, J. Zhao, Q. Jin, Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation, arXiv preprint arXiv:2107.06779 (2021).

[17] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, C. Chen, Hetemotionnet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1047–1056.

[18] J. Li, X. Wang, G. Lv, Z. Zeng, Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation, Neurocomputing (2023) 126427.

[19] J. Li, X. Wang, G. Lv, Z. Zeng, Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition, IEEE Transactions on Multimedia (2023).

[20] J. Li, X. Wang, G. Lv, Z. Zeng, Ga2mif: Graph and attention based two-stage multi-source information fusion for conversational emotion detection, IEEE Transactions on Affective Computing (2023).

[21] W. Shen, J. Chen, X. Quan, Z. Xie, Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 13789–13797.

[22] W. Shen, S. Wu, Y. Yang, X. Quan, Directed acyclic graph network for conversational emotion recognition, arXiv preprint arXiv:2105.12907 (2021).

[23] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[24] 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, volume 2018, NIH Public Access, 2018, p. 2122.

[25] 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.

[26] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, arXiv preprint arXiv:1707.07250 (2017).

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[29] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[31] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, D. Kislyuk, Toward transformer-based object detection, arXiv preprint arXiv:2012.09958 (2020).

[32] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Advances in neural information processing systems 32 (2019).

[33] J. Lin, A. Yang, Y. Zhang, J. Liu, J. Zhou, H. Yang, Interbert: Vision-and-language interaction for multi-modal pretraining, arXiv preprint arXiv:2003.13198 (2020).

[34] Q. Li, Z. Han, X.-M. Wu, Deeper insights into graph convolutional networks for semi-supervised learning, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[35] G. Li, M. Müller, G. Qian, I. C. D. Perez, A. Abualshour, A. K. Thabet, B. Ghanem, Deepgcns: Making gcns go as deep as cnns, IEEE transactions on pattern analysis and machine intelligence (2021).

[36] 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, Language resources and evaluation 42 (2008) 335–359.

[37] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, arXiv preprint arXiv:1810.02508 (2018).

[38] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first

challenge, Speech communication 53 (2011) 1062–1087.
[39] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882 (2014).
[40] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.