Speaker-Aware Interactive Graph Attention Network for Emotion Recognition in Conversation

ZHAOHONG JIA, Anhui University, China
YUNWEI SHI, Anhui University, Hefei Comprehensive National Science Center, China
WEIFENG LIU, Anhui University, China
ZHENHUA HUANG, Anhui University, China
XIAO SUN, Hefei University of Technology Hefei Comprehensive National Science Center, China

Recently, **E**motion **R**ecognition in **C**onversation (ERC) has attracted much attention and has become a hot topic in the field of natural language processing. Conversation is conducted in chronological order; current utterance is more likely influenced by nearby utterances. At the same time, speaker dependency also plays a core role in the conversation dynamic. The combined effect of the sequence-aware information and the speaker-aware information makes the emotion's dynamic change. However, past works used simple information fusion methods to model the two kinds of information but ignored their interactive influence. Thus, we propose a novel method entitled SIGAT (**S**peaker-aware **I**nteractive **G**raph **A**ttention Ne**t**work) to solve the problem. The core module is a mutual interactive module in which a dual-connection (self-connection and interact-connection) graph attention network is constructed. The advantage of SIGAT is modeling the speaker-aware and sequence-aware information in a unified graph and updating them simultaneously. In this way, we model the interactive influence of them and obtain the final representations, which have richer contextual clues. Experimental results on the four public datasets demonstrate that SIGAT outperforms the state-of-the-art models.

CCS Concepts: • **Computing methodologies** → *Discourse, dialogue and pragmatics;*

Additional Key Words and Phrases: Emotion recognition in conversation, text classification, natural language processing

This work is supported by the National Key Research Development Program of China (grant no. 2022YFC3803202), Major Project of Anhui Province (grant no. 202203a05020011), Anhui Province Key Research and Development Program (grant no. 202304a05020068) and General Programmer of the National Natural Science Foundation of China (grant no. 62376084). This research is also supported by the National Natural Science Foundation of China (grant no. 71971002) and by the Major Projects of Science and Technology in Anhui Province (grant no. 202003a060-20016).

Authors' address: Z. Jia, W. Liu, and Z. Huang, Anhui University, School of Computer Science and Technology, HeFei, Anhui, 230601, China; e-mails: zhjia@mail.ustc.edu.cn, {qishengmoming, zhhuangscut}@gmail.com; Y. Shi, Anhui University, School of Computer Science and Technology, Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, HeFei, Anhui, 230601 China; e-mail: 1826719913@qq.com; X. Sun (Corresponding author), Hefei University of Technology, School of Computer Science and Information Engineering, Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, HeFei, Anhui, 230601, China; e-mail: sunx@iai.ustc.edu.cn.

2375-4699/2023/12-ART256 $15.00
https://doi.org/10.1145/3627806

ACM Reference format:
Zhaohong Jia, Yunwei Shi, Weifeng Liu, Zhenhua Huang, and Xiao Sun. 2023. Speaker-Aware Interactive Graph Attention Network for Emotion Recognition in Conversation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22, 12, Article 256 (December 2023), 18 pages.
https://doi.org/10.1145/3627806

1 INTRODUCTION

In recent years, with the application of emotion recognition in conversation in many fields such as opinion mining and social media analysis [26, 32], how to accurately judge the emotion in the conversation has attracted continuous attention in academia and business circles. People can express their emotions through three modalities of information: audio, video, and text. Text information usually contains more emotion clues than video or audio [25, 29]. Therefore, we use text information to identify the emotion of utterances.

There are two key factors that affect emotion recognition in conversation. One is sequence-aware information and the other is speaker-aware information. As shown in Figure 1(a), the current utterance is more affected by the nearby utterances than the distant utterances. This kind of performance fits in with the conversation conducted in chronological order. In addition, speaker-aware information also plays a core role. There are two important dependencies in the emotional dynamics of conversation: intra- and inter-speaker dependency [6]. Intra-speaker dependency, also known as *emotional inertia*, refers to the influence of the speaker on oneself in the conversation. Inter-speaker dependency refers to the emotional influences caused by other speakers in the conversation. These two kinds of information influence each other, leading to the dynamic change of emotion. Thus, it is critical to model the interactive influence of these types of information.

Some past works utilize simple information fusion methods to use the two kinds of information. As shown in Figure 1(b), CoMPM [15] directly adds these two kinds of information to obtain context representation. Similarly, DialogueGCN [5] concates the two kinds of information for the final classification. However, none of the models considers the interactive influence between the two kinds of information, resulting in insufficient context clues in the final representation. EmotionFlow [38] considers the relationship between these two kinds of information by simultaneously decreasing loss. However, it is difficult to clearly control information transfer of the two kinds of information.

To solve these problems, we use an explicit way to establish graph connection to facilitate interaction of the two kinds of information. In this way, they can complement each other and update simultaneously.

In this article, we propose a novel **S**peaker-aware **I**nteractive **G**raph **A**ttention Ne**t**work (SIGAT) to model the mutual interaction of the speaker-aware and sequence-aware information. The advantage of SIGAT is modeling the speaker-aware and sequence-aware information into a unified graph to model their interaction and update them simultaneously. There are three modules in SIGAT: the speaker perception (SP) module, the sequence encoder (SE) module, and the mutual interaction (MI) module. In the SP module, we introduce a speaker-based dynamically updated mechanism based on the speaker-aware mask. The feature matrix will be dynamically updated by the speaker identity of the current aggregated node to perceive the intra- and inter-speaker dependencies. In the updated matrix, the utterance features spoken by the same speaker and different speakers of the current aggregated node will be given different trainable parameters. In addition, to enhance the multi-speaker information, we inject the speaker embedding, which is encoded by the speaker identity to utterance features. In the sequence encoder module, we design a sequence-aware mask to encode the sequence information of utterances. The position weight is calculated as the

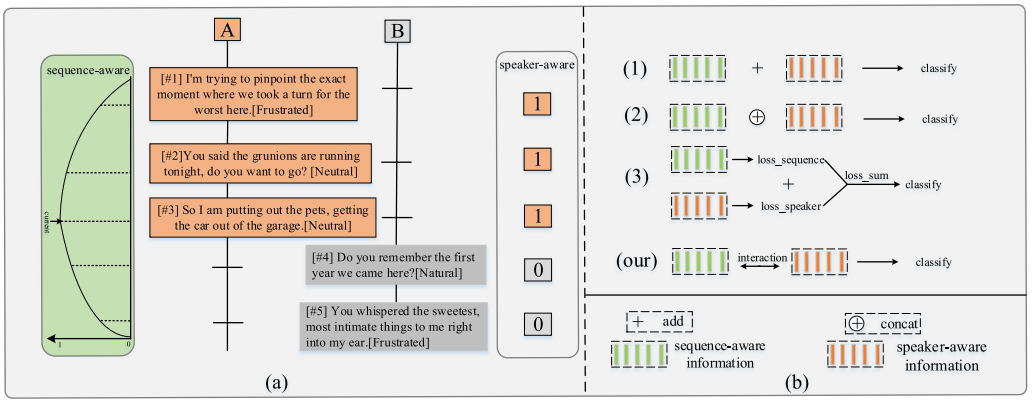


Fig. 1. (a) A conversation segment from the IEMOCAP dataset. Speaker A and Speaker B alternating communicate in the conversation. The current utterance is more affected by the nearby utterances than the distant utterances. Under the influence of speaker dependency, the emotion of utterances is changed dynamically. (b) Different ways of fusing the speaker-aware and sequence-aware information.

sequence-aware mask according to the distance between different utterances. The closer the distance, the greater the position weight of the utterances. Finally, in the mutual interaction module, a dual-connection (self-connection and interact-connection) graph is introduced to enhance the speaker-aware information and sequence-aware information in a mutual way. Through interact-connection, we explicitly established the interaction between these two kinds of information. Through intra-connection, the characteristics of information itself are fully explored. In this way, we model the interactive influence of these two kinds of information. The updated representations, which have richer contextual clues, are used for the final classification. We conduct extensive experiments on four public datasets to verify the effect of our model. The experimental results show that our model outperforms the comparison methods. The contributions of this article are as follows:

— We design a mutual interactive graph network in which the dual-connection is constructed, achieving a model of the interactive influence of speaker-aware information and sequence-aware information.
— We introduce the speaker dynamically updated mechanism based on the speaker-aware mask to perceive the speaker dependency in the aggregating process. In addition, we calculate the position weight by the relative distance of the utterances as the sequence-aware mask.
— We verify the effectiveness of our model on the IEMOCAP, MELD, EmoryNLP, and DailyDialog datasets. Extensive experiments demonstrate the superiority of our model.

2 RELATED WORK

2.1 Emotion Recognition in Conversation (ERC)

In recent years, with the continuous attention to the ERC task, many models have been proposed. These models can be divided into two categories: recurrence-based models and graph-based models.

Recurrence-based Models bc-LSTM [30] uses content-dependent long short-term memory (LSTM) with an attention mechanism to capture contextual information. HIGRU [12] obtains contextual information through the attention mechanism and hierarchical gated recurrent unit

(GRU). BIERU [18] introduces an emotional recurrent unit to obtain contextual information. ICON [6] adds a global GRU based on CMN [7] to improve the previous model. DialogueRNN [27] uses three GRUs to model dialog dynamics and uses the attention mechanism to capture contextual information. COSMIC [4], which adds commonsense knowledge, has a structure similar to DialogueRNN.

Graph-based Models KET [43] combines external commonsense and transformer to model the context through a multi-layer attention mechanism. TODKAT [44] uses a topic-driven and knowledge-aware transformer to detect the emotion in conversation. CTNET [21] uses the transformer to model the intra- and cross-model interactions. HiTrans [16] uses hierarchical transformers to model contextual information. DialogueGCN [5] connects the utterance by a certain size window and uses the relation graph attention network to aggregate information. MMGCN [8] captures inter-modality dependency by the spectral domain GCN [3]. DialogXL [36] makes corresponding improvements to ERC based on XLNET [41]. LR-GCN [34] uses the multi-head attention mechanism to explore the potential connections between utterances. I-GCN [28] uses incremental graph structure to simulate the dynamic process of the conversation. DAG-ERC [37] models conversation context by a directed acyclic graph (DAG).

2.2 Graph Attention Network

Graph-based models [22] have been widely used in many fields. The origin model is the graph convolutional network (GCN) [14], which uses a fixed adjacency matrix as the edge weight. Based on the origin GCN, there are many improved models. The graph attention network [39] fuses the neighborhood nodes information by attention machine, which can calculate the relevance of the connected nodes [10]. The calculation of GAT can be divided into two steps: calculate the attention coefficient between vertices and aggregate the features according to the calculated attention coefficient. RGAT [9] proposes relational position encodings that provide RGAT with sequential information reflecting the relational graph structure.

3 APPROACH

3.1 Problem Definition

In the ERC task, conversation is defined as $\{u_1, u_2, \ldots, u_N\}$, where N denotes the utterance numbers in a conversation. Each utterance $u_i = \{w_1, w_2, \ldots, w_n\}$ is formed by n words. The utterance's speaker identity is denoted by a function $s(\cdot)$. For example, $s(u_i) \in \widetilde{\mathcal{S}}$ denotes the speaker of u_i and $\widetilde{S}$ denotes the collection of all speaker roles in the ERC datasets. Each utterance has an emotion label $y_i \in E$, where E is the emotion label set. Our task is to predict the emotion label y_i of a given utterance u_i in a conversation by the given context information and the speaker information.

Unlike data structures such as sequences and matrices, graph data structures can represent more complex relationships. They have a good application in conversations. Usually, a graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where V represents the set of vertices and E represents the set of edges. For the two vertices i and j of an edge, i and j are adjacent nodes to each other. For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the adjacency matrix is represented as $A \in \mathbb{R}^{V \times V}$, where A_{ij} represents the edge between vertex i and vertex j.

3.2 Utterance Feature Extraction

Following Ghosal et al. [4], we use the utterance feature extracted by RoBERTa-Large [23]. For each utterance u_i, we prepend a special token $[CLS]$ to its tokens, making the input a form of $[CLS], w_1, w_2, \ldots, w_n$. Then, we use the $[CLS]$'s pooled embedding at the last layer as the feature vector of u_i.

3.3 Graph Attention Network

The Graph Attention Network (GAT) calculates the neighborhood's information by attention mechanism. In general, after getting the input vector $\overline{h}^0 = \{\overline{h}_1^0, \ldots, \overline{h}_N^0\}, \overline{h}_n^0 \in \mathbb{R}^F$, the GAT updates the node representation by layers. Weight coefficient α_{ij}^l calculates the correlation between $\overline{h}_i^l$ and $\overline{h}_j^l$ through the attention mechanism:

$$\alpha_{ij}^l = \frac{\exp(\mathcal{F}(\overline{h}_i^l, \overline{h}_j^l))}{\sum_{j \in \mathcal{N}_i} \exp(\mathcal{F}(\overline{h}_i^l, \overline{h}_j^l))}, \tag{1}$$

$$\mathcal{F}\left(\overline{h}_i^l, \overline{h}_j^l\right) = \text{LeakyReLU}\left(\mathbf{a}^\top \left[W_{\overline{h}} \overline{h}_i^l \| W_{\overline{h}} \overline{h}_j^l\right]\right), \tag{2}$$

where l is the current graph layer, $\mathcal{N}_i$ is the neighbor nodes of i in the graph, *LeakyReLU* is an activation function, $W_{\overline{h}} \in \mathbb{R}^{F \times F'}$ is the trainable weight matrix, $\mathbf{a} \in \mathbb{R}^{2 \times F'}$ is also a trainable weight matrix, and F and F' are the input and output dimensions. In addition, the GAT extends the multi-head attention mechanism as follows:

$$\overline{h}_i^{l+1} = \Big\|_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k W_{\overline{h}}^k \overline{h}_j^l\right), \tag{3}$$

where K is the number of attention heads, $W_{\overline{h}}^k$ is the trainable weight matrix, α_{ij}^k is the attention weight at k head, and $||$ is the concatenation operation. Finally, we employ the averaging operation to change the connected dimension to output dimension.

3.4 Speaker-Aware Interactive Graph Attention Network

In this section, we introduce our **S**peaker-aware **i**nteractive **g**raph **a**ttention ne**t**work (SIGAT). SIGAT contains three core modules: the speaker perception module, the sequence encoder module and the mutual interaction module corresponding to Sections 3.4.1, 3.4.2, and 3.4.3, respectively. The framework of SIGAT is shown in Figure 2.

3.4.1 Speaker Perception Module. In this module, we first utilize the graph structure to connect the utterances with a fixed size window. In the aggregating process, we introduce the speaker-based dynamically updated mechanism to update the feature matrix based on the speaker-aware mask. In this way, our model can better perceive the inter- and intra-speaker dependency. In addition, speaker embedding is injected into the utterance features to enhance the multi-speaker information.

Graph Structure: We utilize the graph structure to connect the utterances by the fixed size window. The graph can be denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{u_1, u_2, \ldots, u_N\}$ and the edge $(i, j) \in \mathcal{E}$ represents the utterance connection.

Nodes: We take the utterance feature vector extracted by RoBERTa as the graph node features. The node features can be represented as $h = \{h_1, \ldots, h_N\}, h_i \in \mathbb{R}^d$. Each utterance in the conversation is represented as a vertex $v_i \in \mathcal{V}$ in $\mathcal{G}$.

Edges: In our article, we introduce the past context window and future context window with the size w to construct the graph. Each node h_i has an edge with the near w utterances of the past: $h_{i-1}, h_{i-2}, \ldots h_{i-w}$, w utterances of the future: $h_{i+1}, h_{i+2}, \ldots h_{i+w}$, and itself: h_i. In this way, we fully connected the vertices in a certain size window.

Speaker Embedding: Simplifying intra-speaker and inter-speaker dependency to a binary version [17] is useful in two-speaker conversation but has some shortcomings in multi-speaker

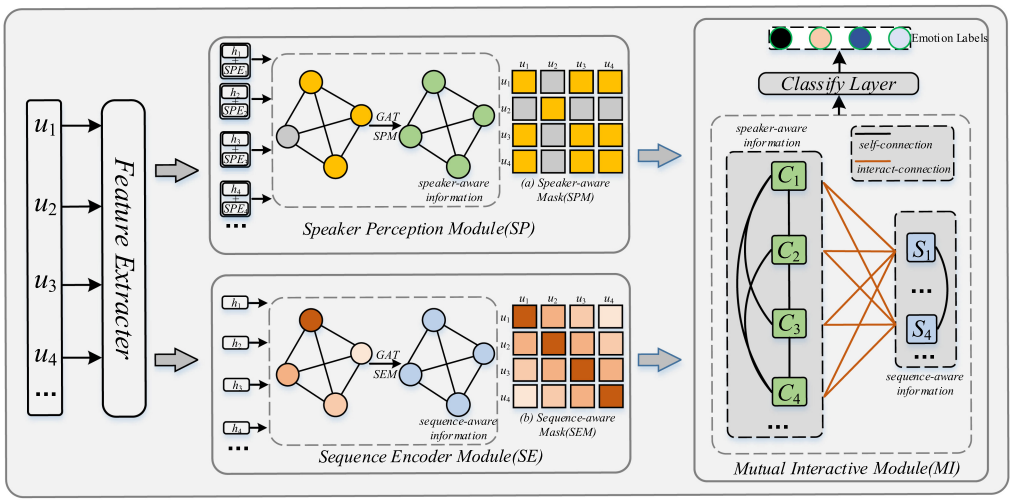


Fig. 2. The framework of our SIGAT. After the feature extraction, we model the speaker-aware and sequence-aware information via the speaker perception module and sequence encoder module. Then, the mutual interaction module is introduced to facilitate interaction between them. Finally, the updated features are used for classification.

conversation [37]. Thus, we inject the speaker embedding to enhance the speaker information in multi-speaker conversation. We first obtain the original speaker embedding $speaker_i$ by randomly initializing the speaker identity $s(u_i)$. After a linear transformation, we get the speaker embedding vector SPE_i of u_i. The speaker embedding vector can then be leveraged to attach the multi-speaker information.

$$SPE_i = W_s * speaker_i + b_s, \tag{4}$$

$$\hat{h}_i = h_i + SPE_i. \tag{5}$$

After obtaining the speaker embedding vector, it is added to h to get $\hat{h} = \{\hat{h}_1, \ldots, \hat{h}_N\}, \hat{h}_i \in \mathbb{R}^d$.

Weight coefficient $\hat{\alpha}_{ij}$ calculates the correlation between $\hat{h}_i$ and $\hat{h}_j$ through attention mechanism:

$$\hat{\alpha}_{ij} = \frac{\exp(\mathcal{F}(\hat{h}_i, \hat{h}_j))}{\sum_{j \in \mathcal{S}_i} \exp(\mathcal{F}(\hat{h}_i, \hat{h}_j))}, \tag{6}$$

$$\mathcal{F}(\hat{h}_i, \hat{h}_j) = \text{LeakyReLU}\left(\mathbf{a}^\top [W_h \hat{h}_i \| W_h \hat{h}_j]\right), \tag{7}$$

where $\mathcal{S}_i$ is the neighbor nodes of i, $\mathbf{a} \in \mathbb{R}^{2\times d}$ is the trainable weight matrix, and d is the hidden dimension.

Speaker-Based Dynamically Update:

Speaker-Aware Mask (SPM): There are two important dependencies in the emotional dynamics of conversation: intra- and inter-speaker dependency [5]. Intra-speaker dependency, also known as *emotional inertia*, refers to the influence of the speaker on oneself in the conversation. Inter-speaker dependency refers to the emotional influences caused by other speakers in the conversation. In order to model the speaker dependency, we design a speaker-aware mask as shown in Figure 3(a). In this mask, if u_i and u_j are said by the same speaker, SPM_{ij} will be marked as 1.

$$SPM_{i,j} = \begin{cases} 1, & if \quad s(u_i) = s(u_j); \\ 0, & otherwise. \end{cases} \tag{8}$$

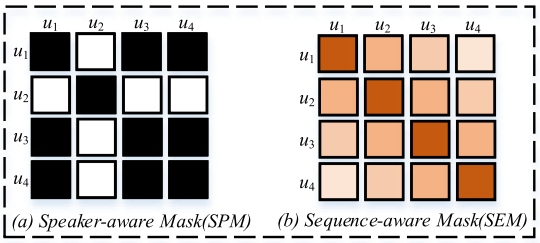


Fig. 3. An example of the speaker-aware mask matrix SPM and the sequence-aware mask SEM. u_1, u_3, u_4 are from the same speaker. In (a), $u_{ij} = 1$ colored black means that u_i and u_j are spoken by the same speaker and $u_{ij} = 0$ colored white means that u_i and u_j are spoken by different speakers. In (b), the position weight is calculated as the sequence-aware mask according to the distance between u_i and u_j. The longer distance between u_i and u_j, the less weight is given to u_j.

After calculating the weight coefficient, we update the feature matrix by the **s**peaker-**b**ased **d**ynamically **u**pdated mechanism (SBDU) inspired by R-GCN [35] to perceive the intra- and inter-speaker dependency in the aggregating process:

$$H_{[i]} = \hat{h} * W_w \odot SPM_i + \hat{h} * W_k \odot (I - SPM_i), \tag{9}$$

where $\odot$ denotes the element-wise multiplication, $SPM \in \mathbb{R}^{N \times N}$ is the speaker mask matrix described in Figure 3, $SPM_{ij} = 1$ means u_i and u_j are said by the same speaker, $SPM_{ij} = 0$ means u_i and u_j are said by different speakers, and SPM_i represents the relationship between the speaker of u_i and the speakers of other utterances. After broadcasting, the dimension of SPM_i changes from 1 to d. $\hat{h} \in \mathbb{R}^{N \times d}$ is the node feature matrix, W_w and $W_k \in \mathbb{R}^{d \times d}$ are two trainable matrices, $I \in \mathbb{R}^{N \times d}$ is a matrix in which all elements are 1, and $H_{[i]} \in \mathbb{R}^{N \times d}$ is the speaker-aware updated matrix of u_i.

After getting the updated feature matrix $H_{[i]}$, we aggregate the information of node j around node i by using the weight coefficient $\hat{\alpha}_{ij}$ calculated before:

$$C_i = \Big\Vert_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{S}_i} \hat{\alpha}_{ij}^k [H_{[i]}]_j\right), \tag{10}$$

where K is the number of attention heads and $||$ is the concatenation operation.

The process of getting the speaker-aware representation can be simplified by the following formula:

$$C = GAT_{speaker}(SPM, \hat{h}), \tag{11}$$

where $\hat{h} = \{\hat{h}_1, \ldots, \hat{h}_N\}, \hat{h}_i$ is the features that added speaker embedding, SPM is the sequence-aware mask, and $C = \{C_1, \ldots, C_N\}$ is the speaker-aware contextual representation.

3.4.2 Sequence Encoder Module. We find that the emotion of the current utterance is greatly affected by the surrounding utterances and that this influence decreases with the increase of distance. Therefore, we calculate the position weight as the sequence-aware mask according to the distance between u_i and u_j as follows.

Sequence-aware Mask (SEM):

$$pw_{ij} = \beta - \lambda(j - i)^2, \tag{12}$$

$$SEM_{ij} = sigmoid(pw_{ij}), \tag{13}$$

where λ and β are hyperparameters, which are set to 6 and 0.2, respectively. Then, the sequence-aware mask will be mutual with the weight coefficient matrix α in the sequence-aware graph and update the utterances features by the graph attention network. The process of getting the sequence-aware representation can be simplified by the following formula:

$$S = GAT_{sequence}(SEM, h), \tag{14}$$

where $h = \{h_1, \ldots, h_N\}$ is the utterance features, SEM is the sequence-aware mask, and $S = \{s_1, \ldots, s_N\}$ is the sequence-aware contextual representation.

3.4.3 Mutual Interaction Module. The advantage of this module is modeling the speaker-aware and sequence-aware information into a unified interaction graph and updating them simultaneously. When exploring the deep semantics of one kind of information, another kind of information will also be considered as auxiliary information.

There are two kinds of edge connections in the graph: self-connection and interact-connection. The former is the internal connection of the information and the latter is the connection between them. Finally, we get the conversation representations with richer contextual clues by modeling the mutual effect and interactive influence between the two kinds of information.

Graph interaction structure has been shown to be effective on various NLP tasks [2, 24, 33]. We apply the interactive graph to the ERC task and achieve it in the following way.

Nodes: After the speaker perception module and the sequence encoder module, we get the speaker-aware contextual represention $C = \{C_1, \ldots, C_N\}$ and the sequence-aware contextual represention $S = \{S_1, \ldots, S_N\}$. There are total $2N$ nodes in the interactive graph with N nodes about the speaker-aware information and the other N nodes about the sequence-aware information. The node representation of interactive graph $D = \{D_1, \ldots, D_{2N}\}$ can be described as follows: $D = \{C, S\} = \{C_1, \ldots, C_N, S_1, \ldots S_N\}$.

There are two types of edges in the graph.

Self-connection: In order to model the internal semantic of the speaker-aware information or sequence-aware information, we fully connect the speaker-aware contextual representation C and sequence-aware contextual representation S, respectively. In the final adjacency matrix $\mathcal{A} \in \mathbb{R}^{2N\times 2N}$, $\mathcal{A}^c$, $\mathcal{A}^s \in \mathbb{R}^{N\times N}$ are the adjacency matrix of the first connection.

Interact-connection: To realize the interaction of two kinds of information, we have established the connection between the two kinds of information. We fully connect the speaker-aware contextual representation node C_i in C with the sequence-aware contextual representation nodes $S_1, \ldots, S_N$. In the final adjacency matrix $\mathcal{A} \in \mathbb{R}^{2N\times 2N}$, $\mathcal{A}^{cs}$, $\mathcal{A}^{sc} \in \mathbb{R}^{N\times N}$ are the adjacency matrices of the second connection corresponding to nodes C and nodes S, respectively.

By doing this, we model the speaker-aware and sequence-aware information into a unified interaction graph and update them simultaneously. The updated process of the mutual interaction module can be formulated as follows.

For the first N nodes C_i, the aggregation formula is as follows:

$$\tilde{C}_i = \Big\Vert_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{A}_i^c} \alpha_{ij}^k W_s^k C_j + \sum_{j\in\mathcal{A}_i^{cs}} \alpha_{ij}^k W_r^k S_j\right). \tag{15}$$

For the last N nodes S_i, the aggregation formula is as follows:

$$\tilde{S}_i = \Big\Vert_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{A}_i^{sc}} \alpha_{ij}^k W_s^k C_j + \sum_{j\in\mathcal{A}_i^{s}} \alpha_{ij}^k W_r^k S_j\right). \tag{16}$$

Table 1. Data Distribution of the Four Datasets

Dataset	# Conversations			# Utterances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5810		1623
MELD	1038	114	280	9989	1109	2610
EmoryNLP	713	99	85	9934	1344	1328
DailyDialog	11118	1000	1000	87170	8069	7740

3.4.4 Training and Prediction. Finally, we concat the first n nodes and the last n nodes as the emotion representation $e = \{e_1, \ldots, e_N\}$, $e_i = [\tilde{C}_i; \tilde{S}_i] \in \mathbb{R}^{N \times 2d}$. Then, we change the hidden dimension to emotion classes through a linear layer. The formulas are as follows:

$$P_i = \text{Softmax}(W_e e_i + b_e), \tag{17}$$

$$y_i = \text{Argmax}(P_i[k]), \tag{18}$$

We use the cross-entropy function to calculate loss:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{M} \sum_{j=1}^{c(i)} log P_{i,j}[y_{i,j}], \tag{19}$$

where M is the number of all conversations, $c(i)$ is the number of utterances in the i-th conversation, $P_{i,j}$ is the probability of the emotional label of utterance j in conversation i, $y_{i,j}$ is the label of utterance j in conversation i, and θ is the trainable parameters.

4 EXPERIMENTS

4.1 Datasets

We conduct detailed experiments on four public datasets. The specific information of the datasets is shown in Table 1.

IEMOCAP [1]: Multi-modal ERC dataset contains videos of two-way conversations. Each video contains a single dialogue. The conversations in IEMOCAP comes from the performance based on a script by two actors. There are in total six emotion classes: *neutral*, *happiness*, *sadness*, *anger*, *frustrated*, and *excited*.

MELD [31]: A multi-speaker and multi-modal ERC dataset collected from the *Friends* TV series. MELD has more than 1,400 dialogues and 13,000 utterances. There are seven emotion classes: *neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, and *fear*.

EmoryNLP [42]: Similar to MELD, EmoryNLP is also collected from the *Friends* TV series, but varies from MELD in the choice of scenes and emotion labels. However, it is different from MELD in the selection of scene and emotion classes. The emotion classes of this dataset include *neutral*, *sad*, *mad*, *scared*, *powerful*, *peaceful*, and *joyful*.

DailyDialog [19]: The dataset collected from human-written daily communications of English learners. There are seven kinds of emotional labels: *neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, and *fear*. DailyDialog does not have speaker information. We regard the dataset as a two-speaker conversation. The speakers alternate in the conversation.

4.2 Experimental Settings

For all datasets, the utterance features are extracted by RoBERTa with the dimension of 1024. We set the size of the hidden dimension d to 100 and the dropout rate to 0.1. The learning rate is set to 0.0001, α of the activation function *LeakyReLU* is set to 0.2, the epoch is set to 100, and the

multi-head numbers M are set to 4, 8, 10, and 8. The window size numbers w are set to 8, 5, 5, and 5. The batch sizes are set to 8, 64, 16, and 16 for the four datasets, respectively.

4.3 Evaluation Metrics

For the DailyDialog dataset, following Ghosal et al. [4], we calculate the micro-averaged F1 (Micro F1) excluding *neutral* labels. Due to the different proportions of emotions in the datasets, weighted F1 scores (W-Avg F1) can better reflect the emotion classification ability of the model. We first calculate the F1 score of each emotion category. Then, we accumulate and sum according to the proportion of different emotions. For the other datasets, we use the weighted-average F1 score (W-Avg F1) as the evaluation metrics [4, 5]. In addition, following Majumder et al. [27], we use Acc. to calculate the classification accuracy.

4.4 Compared Methods

We compare our model with the following baselines in our article:
TextCNN [13] based on the convolutional neural network is context independent.
bc-LSTM [30] introduces the bidirectional LSTM and attention mechanism to model the contextual information.
CMN [7] models the utterances spoken by different speakers respectively, and uses the attention mechanism to fuse context information. ICON [6] adds a global GRU based on CMN to improve the previous model.
DialogueRNN [27] uses three RNNs to model dialog dynamics, which include the speaker, the context, and the emotion of the consecutive utterances and uses the attention mechanism to capture contextual information. DialogueRNN achieves competitive performance as a strong baseline.
AGHMN [11] uses the BiGRU fusion layer to model the correlation of historical utterances. In addition, it uses the attention mechanism to update the internal state of GRU.
A-DMN [40] models self and inter-speaker dependency respectively and then combines these two types of information to update the memory.
BIERU [18] introduces a compact and fast emotional recurrent unit (ERU) to obtain contextual information. BiERU uses a generalized neural tensor block (GNTB) to fuse the context features and a two-channel classifier to perform the emotion classification.
DialogueGCN [5] connects utterances in a certain window size and uses the relation graph attention network to aggregation information. The model uses different edge types to model the inter- and intra-speaker dependency.
TRMSM [17] simplifies the speaker dependency into a binary version and designs three masks for three transformers to model the context information, the intra-speaker information, and the inter-speaker information, respectively. Finally, these types of information are fused through different fusion mechanisms.
COSMIC [4] proposes a new framework that introduces types of different commonsense, such as mental states, events, and causal relations.
DAG-ERC [37] models conversation by a DAG and uses the contextual information unit to enhance the information of historical context. DAG-ERC achieves superior performance as a strong baseline.
CoMPM [15] extracts the pre-trained memory using the pre-trained language model. Then, CoMPM combines the speaker's pre-trained memory with the context model to get the final results.
EmotionFlow [38] designs an end-to-end ERC model, which extracts emotion vectors through the Emoformer structure and obtains the emotion classification results from a context analysis model.

Table 2. The Results of Different Models on the Four ERC Datasets

Model	IEMOCAP W-Avg F1	MELD W-Avg F1	EmoryNLP W-Avg F1	DailyDialog Micro F1
TextCNN [13]	52.04	55.02	32.59	50.32
CMN [43]	56.13	54.50	—	—
DialogueRNN [27]	62.75	57.03	31.70	55.95
AGHMN [11]	63.50	60.30	—	—
A-DMN [40]	64.30	60.50	—	—
BIERU [18]	64.24	60.70	—	—
DialogueGCN [5]	64.18	58.10	31.65	51.25
RoBERTa DialogueGCN [5]	64.91	63.02	38.10	57.52
COSMIC [4]	65.28	65.21	38.11	58.48
DAG-ERC [37]	68.03	63.65	39.02	59.33
EmotionFlow[38]	—	65.08	—	—
CoMPM [15]	69.46	65.77	38.93	59.02
EmoCaps[20]	69.49	63.51	—	—
OUR	**70.17**	**66.18**	**39.95**	**59.92**

EmoCaps [20] proposes a new structure named Emoformer to extract multi-modal emotion vectors from different modalities and fuses them with a sentence vector to be an emotion capsule.

5 RESULTS AND ANALYSIS

5.1 Overall Performance

In Table 2, we list the experimental results of SIGAT and other baseline models.

IEMOCAP and DailyDialog are ERC datasets that have two speakers in the conversation. On the IEMOCAP dataset, our model achieves a new state-of-the-art score of 70.17% on the W-Avg F1. It shows that modeling the interactive influence of the speaker-aware and sequence-aware information is useful in the ERC task. DailyDialog has a large number of *neutral* labels, which affects the final F1 score. Therefore, we remove them in the final calculation following Ghosal et al. [4]. On this dataset, our model achieves 59.92% on the Micro F1 score. MELD and EmoryNLP are multi-speaker ERC datasets. On the datasets, our model achieves 66.18%, 39.95% on the W-Avg F1 score.

BIERU introduces a new recurrence unit ERU to extract context information. However, speaker-aware information is ignored in BiERU. Our SIGAT outperforms BiERU by 5.93% on the IEMOCAP datasets. COSMIC proposes a new framework that introduces different types of commonsense, such as mental states, events, and causal relations based on DialogueRNN. However, RNN cannot process long-distance information well. The multi-head attention in a graph attention network has a stronger ability to explore the semantic relation and aggregate context information. Compared with COSMIC, SIGAT improves by 4.89%, 0.97%, 1.84%, and 1.44%, respectively. Compared with DAG-ERC, SIGAT improves by 2.14%, 2.53%, 0.93%, and 0.59%, respectively. It shows that SIGAT can model the conversation context, which has richer contextual clues. EmoCaps proposes a new structure named Emoformer to extract multi-modal emotion vectors from different modalities. In this article, we adopt the experimental results of EmoCaps on text modality. Compared with EmoCaps, SIGAT improves by 0.68% and 2.67%. CoMPM adds the speaker's pre-trained memory with the sequence context to get the final results. DialogGCN concates the speaker-aware information with sequence-aware information. However, none of these models considers its interaction influence. SIGAT considers the interactive influence of them and obtains the final representations

Table 3. Comparison of Results of Specific Emotion Categories on the IEMOCAP Dataset

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		W-Avg	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TextCNN	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
bc-LSTM	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
ICON	22.22	29.91	58.78	64.57	62.76	57.38	64.71	63.04	58.86	63.42	**67.19**	60.81	59.09	58.54
DialogueRNN	25.69	33.18	75.10	78.80	58.59	59.21	64.71	65.28	80.27	71.86	61.15	58.91	63.40	62.75
AGHMN	48.30	52.10	68.30	73.30	61.60	58.40	57.50	61.90	68.10	69.70	67.10	62.30	63.50	63.50
DialogueGCN	40.62	42.75	**89.14**	**84.54**	61.92	63.54	67.53	64.19	65.46	63.08	64.18	**66.99**	65.25	64.18
A-DMN	43.10	50.60	69.40	76.80	63.00	62.90	63.50	56.50	**88.30**	**77.90**	53.30	55.70	64.60	64.30
TRMSM	43.36	50.22	81.23	75.82	66.11	64.15	60.39	60.97	77.46	72.70	62.16	63.45	65.34	65.74
BIERU	49.81	32.75	81.26	82.37	65.00	60.45	67.86	65.39	63.14	73.29	59.77	60.68	65.35	64.24
OUR	**59.03**	**57.24**	81.63	81.30	**70.57**	**69.04**	**71.18**	**66.67**	74.25	76.65	62.73	65.57	**70.12**	**70.17**

Table 4. Comparison of Results of Specific Emotion Categories on the MELD Dataset

Methods	MELD							
	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	W-Avg F1
TextCNN	34.5	8.2	3.7	49.4	74.9	21.1	45.5	55.0
CMN	44.7	0.0	0.0	44.7	74.3	23.4	47.2	55.5
ICON	44.8	0.0	0.0	50.2	73.6	23.2	50.0	56.3
bc-LSTM	43.4	23.7	9.4	54.5	76.7	24.3	51.0	59.3
DialogueRNN	43.7	7.9	11.7	54.4	77.4	34.6	52.5	60.3
A-DMN	41.0	3.5	8.6	57.4	78.9	24.9	55.4	60.5
TRMSM	46.0	28.6	20.4	58.7	77.6	32.9	57.3	62.4
OUR	**53.4**	**31.2**	**22.5**	**64.6**	**79.4**	**41.5**	**59.4**	**66.2**

that have richer contextual clues. On the IEMOCAP dataset, SIGAT improves by 0.71% and 5.26%, respectively. EmotionFlow considers the relationship between these two kinds of information by simultaneously decreasing losses. It is difficult to clearly control the information transfer of the two kinds of information. SIGAT establishes the explicit connection to control information transfer. Compared with EmotionFlow, SIGAT improves by 1.1% on the MELD dataset.

Then, we will analyze our model and compare it with other models on specific emotional categories on the IEMOCAP and MELD datasets.

There are 6 emotion labels on IEMOCAP: *happy*, *sad*, *neutral*, *angry*, *excited*, and *frustrated*. As shown in Table 3, compared with other models, our model has a certain amount of improvement on *anger*, *frustrated*, and *neutral*. On *neutral*, compared with DialogueGCN, SIGAT improves by 8.65% and 5.5% on Acc. and F1 scores, respectively. On *frustrated*, compared with BIERU-gc, our model improves by 2.96% and 4.89% on Acc. and F1 scores, respectively. As shown in Table 4, our model performs better than the comparison models on majority classification on the MELD dataset. Most of the comparison models perform poorly in emotions that are difficult to distinguish, such as *disgust* and *fear*, and our model has been greatly improved. Compared with DialogueRNN, SIGAT improves by 9.7%, 23.3%, 10.8%, 10.2%, 2%, 6.9%, and 6.9% on the seven emotional categories, respectively.

5.2 Ablation Study

In this section, we will analyze the effects of various modules and mechanisms of our model. The results are shown in Tables 5 and 6.

First, we will analyze the effect of different modules. As shown in Table 5, when the speaker perception module and the sequence encoder module are worked separately, our model achieves

Table 5. The Performance (W-Avg F1) on the IEMOCAP and MELD Datasets with Different Modules of SIGAT

Method	IEMOCAP	MELD
only the speaker perception (SP) module	68.34	64.94
only the sequence encoder (SE) module	67.55	65.40
w/o mutual interaction (MI) module	68.63	64.90
w/o self-connection	69.26	65.46
w/o interact-connection	69.09	65.58
OUR	**70.17**	**66.18**

Table 6. The Performance (W-Avg F1) on the IEMOCAP and EmoryNLP Datasets with Different Mechanisms (SBDU and SEMB) of SIGAT

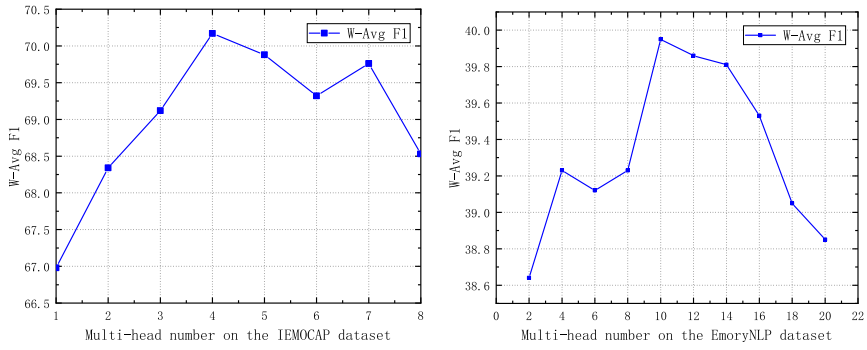
SBDU	SEMB	IEMOCAP	EmoryNLP
✓	✓	70.17	39.95
✗	✓	68.92	39.43
✓	✗	69.25	38.73
✗	✗	68.36	38.55

68.34% and 67.55% for the W-Avg F1 on the IEMOCAP dataset, respectively. After the interaction of the speaker-aware information and the sequence-aware information by the mutual interaction module, the W-Avg F1 is improved by 1.83% and 2.62%, respectively. Next, we will analyze the effort of the two connections in the mutual interaction module. When the self-connection is removed, the W-Avg F1 is reduced by 0.91% and 0.72% on the IEMOCAP and MELD, respectively. When the interact-connection is removed, the W-Avg F1 is reduced by 1.08% and 0.6%, respectively. The best results are obtained when the two connections exist at the same time. When the mutual interactive module is removed, the speaker-aware information and sequence-aware information will be simply connected. Under these circumstances, the results are reduced by 1.54% and 1.28%, respectively. These phenomena show that model of the integration of the speaker-aware information and sequence-aware information will obtain richer contextual clues and improve performance.

Second, we will analyze the effect of the mechanisms used in the speaker perception module by removing the speaker-based dynamically updated (SBDU) and speaker embedding (SEMB) mechanisms, respectively. IEMOCAP is a dataset of two-speaker conversation and EmoryNLP is a dataset of multi-speaker conversation. In our model, we combine SBDU and SEMB mechanisms to better model the speaker-aware information. As shown in Table 6, the W-Avg F1 drops sharply by 1.25% on the IEMOCAP dataset after removing the SBDU mechanism. The experimental result shows that it is necessary to model the speaker-aware information in the ERC task. When the SEMB mechanism is removed, the result is reduced by 0.92%. EmoryNLP is a multi-speaker conversation dataset that has 9 speakers in total. The results are reduced by 0.52% and 1.22% after removing the SBDU and SEMB mechanisms, respectively. When the two mechanisms are removed at the same time, the result is reduced by 1.4%. The results show that it is not enough to only use the SBDU mechanism on multi-speaker datasets. SIGAT, which combines the SEMB and SBDU, is more useful.

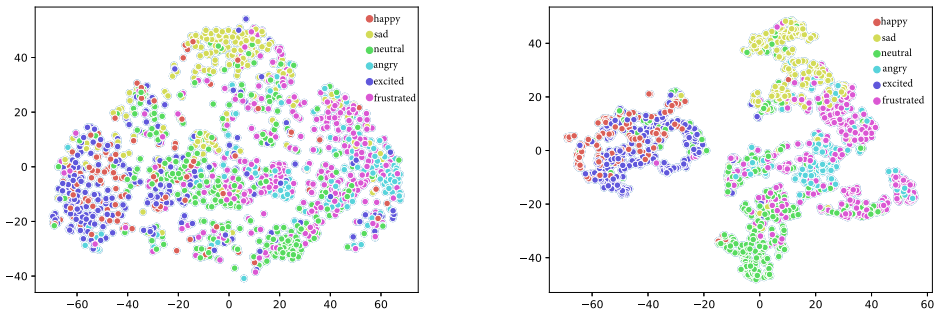
5.3 Analysis on Parameters

In this section, we will analyze the effort of multi-head number M in our model.



(a) M(IEMOCAP) (b) M(EmoryNLP)

Fig. 4. Parameter analysis of the multi-head number (M).



(a) Origin features (b) Features learned by SIGAT

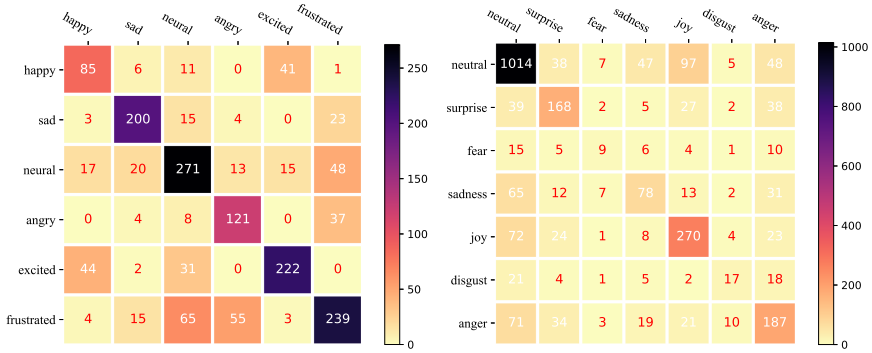
Fig. 5. Results of T-SNE visualization.

In order to stabilize the learning process of self-attention, the graph attention network uses multi-head attention to expand the attention mechanism. We use the multi-head attention mechanism to explore the potential semantic relation of conversation context. As shown in Figure 4, we select the multi-head number from {2, 4, 6, 8, 10, 12, 14, 16, 18, 20} for the EmoryNLP dataset and {1, 2, 3, 4, 5, 6, 7, 8} for the IEMOCAP dataset. On the IEMOCAP dataset, with the multi-head number increasing from 1 to 4, W-Avg F1 is generally in an upward trend. When M=4, SIGAT achieves the best result, 70.17%. After this point, W-Avg F1 begins to decline with the number increasing. On the EmoryNLP dataset, when M=10, SIGAT achieves the best result, 39.95%. After this point, W-Avg F1 begins to decline with the number increasing.

These phenomena show that the increase of M will improve the ability to explore the semantic relation of utterances and model the context. However, continuously increasing the multi-head number will bring redundant information and reduce the final results.

5.4 T-SNE visualization and confusion matrix

In this section, we will analyze the t-SNE visualization results of the model output on the IEMOCAP dataset. Figure 5(a) presents the origin feature distribution learned by RoBERTa. Figure 5(b) presents the feature distribution learned by SIGAT. Compared with the original feature learned by RoBERTa, the features learned by SIGAT are more concentrated. The emotion of *sad*, *excited*, and *neutral* could be effectively clustered by our SIGAT. Our model learns the interactive influence of



(a) Confusion matrix on the IEMOCAP dataset. (b) Confusion matrix on the MELD dataset.

Fig. 6. Confusion matrix.

Table 7. The Case Study Example on the IMEOCAP Dataset

Turn	Speaker	Utterance	Truth	OUR
#48	A	Now Chris you can go on from there, but I don't know what to do. I mean do you know what to do, because I don't.	Neutral	Neutral
#49	A	I don't know why it is. But every time I reach out for something I want, I have to pull back because other people will suffer. My whole bloody life, time after time, after time.	Neutral	Neutral
#50	B	Well you're a considerate fellow, there's nothing wrong with that.	Sad	Sad
#51	A	To hell with that!	Frustrated	Frustrated
#52	B	Have you asked Annie yet to marry you.	Neutral	Sad
#53	B	I want to get this settled first.	Sad	Sad

speaker-aware and sequence-aware information, which makes the final features contain more contextual clues. The features learned by our model could be better used for the final classification. In addition, *excited* and *happy* are similar expressions; the feature distributions of *excited* and *happy* are intermingled.

As shown in the confusion matrix heatmap on the IEMOCAP dataset in Figure 6, our model has some shortcomings in distinguishing similar emotions, especially between *happy* and *excited*. There are many *happy* labels misjudged as *excited* and many *excited* labels are misjudged as *happy*. On another pair, similar emotions *sad* and *frustrated*, our model performs slightly better.

There are only 50 samples of *Fear* and only 68 samples of *Disgust* on the MELD test set. However, there are 1,256 shapes of *neutral* on the test set. Because of this uneven distribution, our model tends to misjudge other emotions as *natural*. On *fear* and *disgust*, the lack of samples leads to the low accuracy of our model.

5.5 Case Study

In this section, we list the prediction results of a fragment on the test dataset on IEMOCAP as shown in Table 7. On the whole, in this conversation segment, our model has a good performance and correctly judges all *neutral* labels. The emotion change of #49 and #51 reflects the combined

effect of the speaker-aware and sequence-aware information in conversation. With the change of time and the influence of speaker A, the emotion of speaker B suddenly changes from *neutral* to *frustrated*. Our model utilizes the mutual interaction graph to explore the interactive influence between these types of information. Finally, we obtain the representations that contain more conversational clues. In the face of the above emotion transfer phenomenon, SIGAT still makes a correct judgment.

6 CONCLUSION

In this article, we propose a model entitled **S**peaker-aware **I**nteractive **G**raph **A**ttention Ne**t**work (SIGAT) that considers the interactive influence between speaker-aware and sequence-aware information. We utilize two mask mechanisms (speaker-aware mask and sequence-aware mask) to perceive these two kinds of information, respectively. In the mutual interaction module, a dual-connection (self-connection and interact-connection) graph is introduced to enhance the speaker-aware information and sequence-aware information in a mutual way. The advantage of this module is modeling the speaker-aware and sequence-aware information into a unified interaction graph and updating them simultaneously. In this way, we obtain the final representations that have more contextual clues. Empirical results on four ERC datasets show that SIGAT is effective in solving the ERC task. Our future work will focus on the application of multi-modal features.

REFERENCES

[1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 4 (2008), 335.

[2] Zi Chai and Xiaojun Wan. 2020. Learning to ask more: Semi-autoregressive sequential question generation under dual-graph interaction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 225–237.

[3] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. 2020. Simple and deep graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, Vol. 119. 1725–1735.

[4] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2470–2481.

[5] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 154–164.

[6] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2594–2604.

[7] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Vol. 1. 2122–2132.

[8] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP (Volume 1: Long Papers)*. 5666–5675.

[9] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. 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)*. 7360–7370.

[10] Yanbin Jiang, Huifang Ma, Yuhang Liu, Zhixin Li, and Liang Chang. 2021. Enhancing social recommendation via two-level graph attentional networks. *Neurocomputing* 449 (2021), 71–84.

[11] Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *The 34th AAAI Conference on Artificial Intelligence, AAAI 2020*. 8002–8009.

[12] Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 397–406.
[13] Y. Kim. 2014. Convolutional neural networks for sentence classification. arXiv:1408.5882.
[14] Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
[15] Joosung Lee and Wooin Lee. 2022. CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*. Association for Computational Linguistics, 5669–5679.
[16] Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4190–4200.
[17] Jiangnan Li, Zheng Lin, Peng Fu, Qingyi Si, and Weiping Wang. 2020. A hierarchical transformer with speaker modeling for emotion recognition in conversation. *arXiv preprint arXiv:2012.14781* (2020).
[18] Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2022. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing* 467 (2022), 73–82.
[19] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 986–995.
[20] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). 1610–1618.
[21] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* (2021), 985–1000.
[22] Hongmin Liu, Zhenzhen Xiao, Bin Fan, Hui Zeng, Yifan Zhang, and Guoquan Jiang. 2021. PrGCN: Probability prediction with graph convolutional network for person re-identification. *Neurocomputing* 423 (2021), 57–70.
[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[24] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 505–514.
[25] H. Ma, J. Wang, L. Qian, and H. Lin. 2021. HAN-ReGRU: Hierarchical attention network with residual gated recurrent unit for emotion recognition in conversation. *Neural Computing and Applications* 33, 3 (2021).
[26] Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Inf. Fusion* 64 (2020), 50–70.
[27] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6818–6825.
[28] Weizhi Nie, Rihao Chang, Minjie Ren, Yuting Su, and Anan Liu. 2021. I-GCN: Incremental graph convolution network for conversation emotion detection. *IEEE Transactions on Multimedia* (2021).
[29] Soujanya Poria, Erik Cambria, and Alexander F. Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). 2539–2544.
[30] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. 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)*. 873–883.
[31] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Rada Mihalcea, Gautam Naik, and Erik Cambria. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL 2019: The 57th Annual Meeting of the Association for Computational Linguistics*. 527–536.
[32] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019a. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access* 7 (2019a), 100943–100953.
[33] Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-GAT: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13709–13717.

[34] Minjie Ren, Xiangdong Huang, Wenhui Li, Dan Song, and Weizhi Nie. 2021. LR-GCN: Latent relation-aware graph convolutional network for conversational emotion recognition. *IEEE Transactions on Multimedia* (2021).
[35] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web — 15th International Conference, ESWC 2018*, Vol. 10843. 593–607.
[36] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. DialogXL: All-in-one XLNet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 13789–13797.
[37] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. 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)*. 1551–1560.
[38] Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022. EmotionFlow: Capture the dialogue level emotion transitions. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022*. IEEE, 8542–8546.
[39] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations, ICLR*.
[40] Songlong Xing, Sijie Mai, and Haifeng Hu. 2020. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing* (2020).
[41] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*. 5753–5763.
[42] Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In *AAAI Workshops*. 44–52.
[43] Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 165–176.
[44] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021 (Volume 1: Long Papers), 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). 1571–1582.

Received 29 March 2022; revised 22 May 2023; accepted 28 September 2023