DualGATs: Dual Graph Attention Networks for Emotion Recognition in Conversations

Duzhen Zhang
Baidu Inc, Beijing, China
zhangduzhen@baidu.com

Feilong Chen
Huawei Inc, Beijing, China
chenfeilong10@huawei.com

Xiuyi Chen*
Baidu Inc, Beijing, China
chenxiuyi01@baidu.com

Abstract

Capturing complex contextual dependencies plays a vital role in Emotion Recognition in Conversations (ERC). Previous studies have predominantly focused on speaker-aware context modeling, overlooking the discourse structure of the conversation. In this paper, we introduce Dual Graph ATtention networks (DualGATs) to concurrently consider the complementary aspects of discourse structure and speaker-aware context, aiming for more precise ERC. Specifically, we devise a Discourse-aware GAT (DisGAT) module to incorporate discourse structural information by analyzing the discourse dependencies between utterances. Additionally, we develop a Speaker-aware GAT (SpkGAT) module to incorporate speaker-aware contextual information by considering the speaker dependencies between utterances. Furthermore, we design an interaction module that facilitates the integration of the DisGAT and SpkGAT modules, enabling the effective interchange of relevant information between the two modules. We extensively evaluate our method on four datasets, and experimental results demonstrate that our proposed DualGATs surpass state-of-the-art baselines on the majority of the datasets.[1]

1 Introduction

With the increasing availability of conversational data on social media platforms (Poria et al., 2019a), Emotion Recognition in Conversations (ERC) has emerged as a popular research topic (Poria et al., 2019b). Its objective is to identify and track the emotional state of each utterance. ERC plays a crucial role in various applications, including opinion mining in social media (Chatterjee et al., 2019b) and the development of empathetic dialogue systems that can analyze user emotional states and generate emotion-aware responses (Zhou et al., 2018; Liu et al., 2021; Peng et al., 2022, 2023).

However, analyzing emotions in conversations poses significant challenges. Unlike emotion recognition in isolated sentences (Seyeditabari et al., 2018), ERC requires careful consideration of contextual dependencies. Previous ERC methods have primarily focused on capturing speaker or temporal dependencies between utterances, making the modeling of speaker-aware context central to these approaches (Majumder et al., 2019).

To incorporate speaker-aware contextual information, numerous methods have been proposed to model conversations as sequences (Poria et al., 2017; Hazarika et al., 2018a,b; Jiao et al., 2019; Hu et al., 2021; Ong et al., 2022) or graphs (Ghosal et al., 2019; Ishiwatari et al., 2020; Shen et al., 2021b; Li et al., 2022). Sequence-based methods capture sequential information by encoding utterances temporally using Recurrent Neural Networks (RNNs). Majumder et al. 2019 designed an independent Gated Recurrent Unit (GRU) (Cho et al., 2014) to track the emotional state of the speaker. However, these sequence-based methods often rely on limited information from nearby utterances to update the current utterance's representation, making it challenging to capture distant contextual information and achieve satisfactory performance. To address this limitation, graph-based methods simultaneously aggregate information from surrounding contextual utterances to update the representation of the current utterance using Graph Neural Networks (GNNs) (Kipf and Welling, 2017). These methods typically treat the conversation as a directed graph, where nodes represent utterances, edges indicate dependency links between pairs of nodes, and edge labels denote the dependency types, such as speaker or temporal relationships.

Despite the remarkable progress made by sequence-based and graph-based methods, there is a need for greater emphasis on explicitly model-

*Corresponding author.
[1] Available at https://github.com/BladeDancer957/DualGATs.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7395–7408

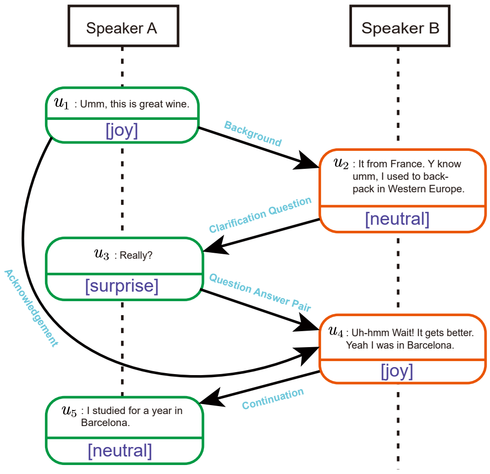

Figure 1: A conversation from MELD dataset (Poria et al., 2019a) with discourse dependencies extracted from a discourse parser (Shi and Huang, 2019).

ing discourse structure within conversations. Discourse structure, which includes discourse dependency links and their types between utterances, offers a straightforward way to capture the essential information flow in a conversation. As illustrated in Figure 1, highly relevant utterances are linked based on discourse dependency types such as Background, Acknowledgement, and Question-Answer Pair. Explicitly incorporating these discourse dependencies in conversations can assist models in capturing significant contextual cues that influence emotions. For instance, let's consider the first and fourth utterances in Figure 1, where there exists a direct discourse dependency link of *Acknowledge* type between utterances 1 and 4. In utterance 1, Speaker A expresses a positive opinion about wine, conveying a sense of *joy* emotion. Speaker B strongly acknowledges this opinion in utterance 4, stating that the wine improves and also experiences a sense of *joy* emotion.

In this paper, we propose a novel method called Dual Graph ATtention networks (DualGATs) that aims to improve the accuracy of ERC by simultaneously considering the complementarity of discourse structure and speaker-aware context. The DualGATs layer comprises three components: Discourse-aware GAT (DisGAT), Speaker-aware GAT (SpkGAT), and an interaction module. The DisGAT module is designed to capture structural-level correlations among the interactive turns explicitly. It propagates the message over the discourse dependency graph obtained from a discourse parser (Shi and Huang, 2019), thereby incorporating discourse structural information. On the other hand, the SpkGAT module is implicitly organized to capture semantic-level correlations among the interactive turns. It conducts message propagation over the speaker dependency graph, constructed based on speaker identities and the relative positions of utterances, enabling the incorporation of speaker-aware contextual information. Furthermore, inspired by previous work (Li et al., 2021b; Zhang et al., 2022), the interaction module leverages mutual cross-attention to integrate the DisGAT and SpkGAT modules, facilitating the exchange of relevant information between the two modules. To enhance the complementarity of the learned representations from the DisGAT and SpkGAT modules and minimize overlap, the interaction module also includes a differential regularizer. This regularizer encourages the two modules to capture different contextual information.

Our contributions can be summarized as follows:

- We propose DualGATs to simultaneously consider the complementarity of discourse structure and speaker-aware context for more precise and accurate ERC.
- We introduce an interaction module to exchange the relevant information between the SpkGAT and DisGAT modules by mutual cross-attention, where a differential regularizer is proposed to induce the two modules to capture different contextual information.
- We conduct extensive experiments on four publicly available ERC datasets. The results of our experiments demonstrate that DualGATs outperform state-of-the-art baselines on most of the tested datasets. Further analyses validate the effectiveness of the critical components in DualGATs.

2 Related Work

2.1 ERC

Recently, due to the proliferation of publicly available conversational datasets (Chen et al., 2019; Chatterjee et al., 2019a), ERC has increasingly become a popular research topic, including the text-modality and multi-modality settings (Zhang et al., 2023; Chen et al., 2023). Here, we specifically focus on the former. Previous studies primarily concentrate on modeling speaker-aware conversational context. Early methods rely on RNNs to encode utterances temporally and track the speaker's state (Jiao et al., 2019; Hu et al., 2021). Notably, BC-LSTM (Poria et al., 2017) employs Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997), while ICON (Hazarika et al., 2018a) and CMN (Hazarika et al., 2018b) utilize GRUs (Cho et al., 2014) and memory networks. DialogueRNN (Majumder et al., 2019) utilizes three GRUs to capture speaker, temporal, and emotional dependencies among utterances. However, these sequence-based methods often rely on limited information from nearby utterances to update the state of the current utterance, which poses challenges in capturing long-range contextual information.

To model the global conversational context, various graph-based methods have emerged (Zhang et al., 2019; Shen et al., 2021a). DialogueGCN (Ghosal et al., 2019) treats each conversation as a fully-connected graph, where nodes represent utterances and edges denote speaker and temporal dependencies between utterances. RGAT (Ishiwatari et al., 2020) introduces relational position encoding to incorporate position information into the GNNs explicitly. DAG-ERC (Shen et al., 2021b) utilizes directed acyclic graphs to model the interaction between speakers and utterances. Additionally, there are several Transformer-based methods (Vaswani et al., 2017) for modeling the conversational context. Since the self-attention module in Transformer can be seen as a fully-connected graph, we consider some Transformer-based approaches as graph-based methods. CoG-BART (Li et al., 2022) employs BART (Lewis et al., 2020) as an utterance encoder and incorporates an auxiliary response generation task to enhance the model's ability to handle contextual information. It also leverages contrastive learning to improve the identification of similar emotions. CoMPM (Lee and Lee, 2021) introduces a pre-trained memory module to consider the linguistic preferences of speakers.

Since humans do not always explicitly express their emotions in their words, there are many methods to incorporate additional general information into the sequence- or graph-based methods to enhance the understanding of implicit emotions. For example, KET (Zhong et al., 2019), KAITML (Zhang et al., 2020), and COSMIC (Ghosal et al., 2020) introduce commonsense knowledge, TODKAT (Zhu et al., 2021) integrates topic information, KI-Net (Xie et al., 2021) leverages sentiment lexicons, DialogueRole (Ong et al., 2022) incorporates utterance role information, SKAIG (Li et al., 2021a) fuses psychological knowledge, and CauAIN (Zhao et al., 2022) includes emotion cause information to enhance ERC.

Despite significant progress, these methods above need to pay more attention to the importance of conversational discourse structure in capturing salient contextual cues that influence emotion. However, due to the complexity of human-human interaction, GNNs (Kipf and Welling, 2017; Yu et al., 2022) directly over the parsed discourse dependency graph like DisGCN (Sun et al., 2021) may not work well as expected on datasets that are not sensitive to discourse structure. Instead of relying solely on discourse structure, we integrate it into our carefully designed DualGATs framework to simultaneously consider discourse structure's and speaker-aware context's complementarity. This integration allows us to achieve more accurate ERC by leveraging the benefits of both aspects.

2.2 Discourse Parsing

Recently, deep sequential models have emerged as practical approaches for conversational discourse parsing (Shi and Huang, 2019; Liu and Chen, 2021). These models have proven their efficacy in various dialogue understanding tasks, such as multi-turn response selection (Jia et al., 2020), as well as dialogue generation tasks, including conversation summarization (Chen and Yang, 2021; Feng et al., 2021). In our work, the discourse structures on which our DisGAT module relies are also parsed using deep sequential models (Shi and Huang, 2019). Leveraging discourse dependencies intuitively enables the model to encode unstructured human conversations better and focus on salient utterances, leading to more accurate predictions.

3 Methodology

We begin by providing a formal definition of the ERC task. A conversation is represented as a sequence of utterances $(u_i, s_i)|i = 1, ..., N$, where each utterance u_i is spoken by speaker s_i, and N denotes the total number of utterances. The objective of the ERC task is to assign an emotion label $y_i \in \mathcal{Y}$, such as joy, sadness, etc., to each utterance u_i in the conversation, where $\mathcal{Y}$ represents the set of possible emotion labels.

The proposed DualGATs consist of three main components: feature extraction, DualGATs layer, and emotion prediction. The overall architecture of our DualGATs is illustrated in Figure 2.

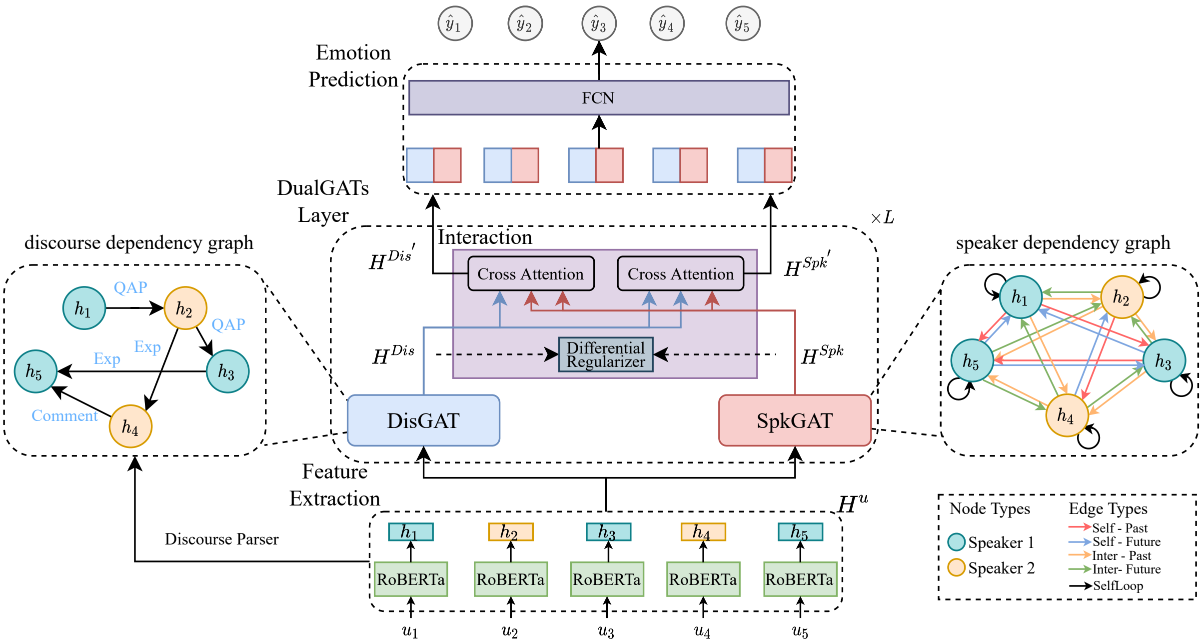

Figure 2: The overall architecture of our DualGATs, encompassing three essential modules: DisGAT, SpkGAT, and Interaction. DisGAT propagates discourse structural information by leveraging discourse dependencies between utterances, while SpkGAT propagates speaker-aware contextual information considering speaker and temporal dependencies. The interaction module initially employs a differential regularizer to ensure that the DisGAT and SpkGAT modules capture distinct contextual information. Subsequently, it utilizes mutual cross-attention to integrate the DisGAT and SpkGAT modules, facilitating the exchange of relevant information between them. In the diagram, the discourse dependency types *Question-Answer Pair* (QAP) and *Explanation* (Exp) are denoted.

3.1 Feature Extraction

In line with previous methods (Ghosal et al., 2020; Shen et al., 2021b), we employ the RoBERTa Large model (Liu et al., 2019) to extract utterance features. The RoBERTa Large model is first fine-tuned for emotion prediction using the transcript of utterances and subsequently utilized as a feature extractor with frozen parameters during the training of DualGATs. Specifically, for the i-th utterance u_i, we prepend a special token "[CLS]" to its tokens, resulting in an input format of $\{[CLS], w_1, ..., w_{n_i}\}$, where n_i denotes the number of tokens in u_i. Subsequently, we extract the output activations from the last layer corresponding to the "[CLS]" token, which serves as the feature representation $\boldsymbol{h}_i \in \mathbb{R}^{d_u}$ of u_i. Here, d_u represents the dimension of the feature representation. Collectively, the feature representations for all utterances are represented as $\boldsymbol{H}^u \in \mathbb{R}^{N \times d_u}$.

3.2 DualGATs Layer

The DualGATs layer efficiently captures both discourse structure and speaker-aware context within a conversation, employing three essential modules: DisGAT, SpkGAT, and Interaction. In this section, we first outline the computation process for each module in the initial layer and then extend it to multiple subsequent layers.

DisGAT The DisGAT module performs message propagation over a discourse dependency graph to integrate discourse structural information. We begin by explaining the construction of the discourse dependency graph, followed by an overview of the inference process employed by the DisGAT module on the constructed graph.

We define the discourse dependency graph of a conversation as $\mathcal{G}^{Dis} = (\boldsymbol{V}^{Dis}, \boldsymbol{E}^{Dis})$, where $\boldsymbol{V}^{Dis}$ represents the set of nodes representing Elementary Discourse Units (EDUs), and $\boldsymbol{E}^{Dis}$ is the adjacency matrix that describes the discourse dependencies between EDUs. In our approach, each utterance in the conversation is treated as an EDU, and we leverage the 16 discourse dependency types outlined in (Asher et al., 2016). These dependency types encompass *Comment*, *Clarification Question*, *Elaboration*, *Acknowledgment*, *Continuation*, *Explanation*, *Conditional*, *Question-Answer Pair*, *Alternation*, *Question-Elaboration*, *Result*, *Background*, *Narration*, *Correction*, *Parallel*, and *Contrast* (We refer to this set of types as R^{Dis}).

Specifically, we first pre-train a discourse parser (Shi and Huang, 2019) on a human-annotated dialogue corpus (Asher et al., 2016), with a 0.78 F1 on link predictions and 0.56 F1 on relation classifications, comparable to the state-of-the-art results. Then, we use this pre-trained parser to predict the discourse dependencies within conversations present in ERC datasets. Consequently, for each conversation, we represent its corresponding discourse dependency graph as $\mathcal{G}^{Dis} = (\boldsymbol{V}^{Dis}, \boldsymbol{E}^{Dis})$. Here, $\boldsymbol{V}^{Dis}[i]$ or v_i^{Dis} represents the node corresponding to utterance u_i, initialized with the corresponding feature representation $\boldsymbol{h}_i$. The edge $\boldsymbol{E}^{Dis}[i][j]$ or $e_{i,j}^{Dis}$ is assigned the dependency type $r^{Dis} \in R^{Dis}$ if a link exists from u_i to u_j with that specific type. This is illustrated in the left part of Figure 2.

Once the discourse dependency graph for the conversation is constructed, we apply the DisGAT module to propagate and aggregate discourse structural information among the graph nodes. The DisGAT module is built upon GAT (Veličković et al., 2018) but includes type coding to account for the dependency types between nodes (utterances). Specifically, for a given node v_i^{Dis}, the DisGAT aggregates the information of its neighboring nodes as follows:

$$\alpha_{ij} = sm_i(LRL(\boldsymbol{a}^T[\boldsymbol{W}\boldsymbol{h}_i \| \boldsymbol{W}\boldsymbol{h}_j \| \boldsymbol{e}_{ij}^{Dis}]))$$
$$\boldsymbol{h}_i^{Dis} = \sum_{j \in \mathcal{N}_i^{Dis}} \alpha_{ij} \boldsymbol{W}\boldsymbol{h}_j \tag{1}$$

where α_{ij} denotes the edge weight from node v_i^{Dis} to its neighbor v_j^{Dis}, *sm* denotes *softmax* function, *LRL* denotes *LeakyReLU* activation function, $\boldsymbol{W}$ and $\boldsymbol{a}$ denote trainable parameters, $\boldsymbol{e}_{ij}^{Dis} \in \mathbb{R}^{|R^{Dis}|}$ denotes the one-hot coding (fixed during model training) corresponding to the discourse dependency type between nodes v_i^{Dis} and v_j^{Dis}, $|R^{Dis}|$ denotes the number of discourse dependency types, $\|$ denotes a concatenation operation, $\mathcal{N}_i^{Dis}$ denotes the neighbours of node v_i^{Dis} in $\mathcal{G}^{Dis}$, $\boldsymbol{h}_i^{Dis} \in \mathbb{R}^{d_h}$ denotes the hidden representation associated with node v_i^{Dis} after DisGAT update, and d_h denotes the dimension of the hidden representation. The updated hidden representation of all nodes is denoted as $\boldsymbol{H}^{Dis} \in \mathbb{R}^{N \times d_h}$.

We summarize the calculation process of the DisGAT in the initial layer as follows:

$$\boldsymbol{H}^{Dis} = \textbf{DisGAT}(\boldsymbol{H}^u, \boldsymbol{E}^{Dis}) \tag{2}$$

SpkGAT The SpkGAT module performs message propagation on a speaker dependency graph to incorporate speaker-aware contextual information. We will first explain the construction of the speaker dependency graph and then introduce the inference process of the SpkGAT on this constructed graph.

We define the speaker dependency graph of a conversation as $\mathcal{G}^{Spk} = (\boldsymbol{V}^{Spk}, \boldsymbol{E}^{Spk})$, where $\boldsymbol{V}^{Spk}[i]$ or v_i^{Spk} represents u_i (the i-th utterance), and its representation is initialized with the corresponding feature representation $\boldsymbol{h}_i$. $\boldsymbol{E}^{Spk}$ is the adjacency matrix that describes the speaker along with temporal dependencies between nodes (utterances). Following the conventions of previous graph-based ERC methods (Ghosal et al., 2019; Ishiwatari et al., 2020), we define five speaker dependency types: *Self-Past*, *Self-Future*, *Inter-Past*, *Inter-Future*, and *SelfLoop* (referred to as set R^{Spk}). Specifically, *Self* represents the influence of the current utterance on other utterances expressed by the same speaker. *Inter* indicates the influence of the current utterance on those expressed by other speakers (excluding the speaker of the current utterance). *Past* and *Future* refer to the relative position of the current utterance and other utterances in the conversation, determining how past utterances influence future utterances and vice versa.[2] *SelfLoop* signifies the self-influence of the current utterance. For any u_i and u_j, $\boldsymbol{E}^{Spk}[i][j]$ or $e_{i,j}^{Spk} = r^{Spk}$ if they satisfy the speaker dependency type $r^{Spk} \in R^{Spk}$ (as depicted in the right part of Figure 2).

After constructing the speaker dependency graph for the conversation, we implement the SpkGAT to propagate and aggregate speaker-aware contextual information across the graph nodes. Similarly, the calculation process of the SpkGAT in the initial layer is summarized as follows:

$$\boldsymbol{H}^{Spk} = \textbf{SpkGAT}(\boldsymbol{H}^u, \boldsymbol{E}^{Spk}) \tag{3}$$

Interaction Module To capture distinct information from the discourse structure and speaker-aware context, we introduce a differential regularizer that encourages divergence between the updated representations of the DisGAT and SpkGAT modules. The regularizer is formulated as follows:

$$\ell_{reg} = \frac{1}{||\boldsymbol{H}^{Dis} - \boldsymbol{H}^{Spk}||_F} \tag{4}$$

[2] Since ERC is viewed as an offline task and future dependencies may help the model fill in some missing information, like the speaker's background, we consider future influence.

where the subscript F denotes the Frobenius norm.

Then, to integrate the DisGAT and SpkGAT modules and effectively exchange relevant information between the two modules, we adopt a mutual cross-attention as a bridge. The computation process is formulated as follows:

$$\begin{aligned} \boldsymbol{A}_1 &= \mathit{softmax}(\boldsymbol{H}^{Dis}\boldsymbol{W}_1(\boldsymbol{H}^{Spk})^T) \\ \boldsymbol{A}_2 &= \mathit{softmax}(\boldsymbol{H}^{Spk}\boldsymbol{W}_2(\boldsymbol{H}^{Dis})^T) \\ \boldsymbol{H}^{Dis'}, \boldsymbol{H}^{Spk'} &= \boldsymbol{A}_1\boldsymbol{H}^{Spk}, \boldsymbol{A}_2\boldsymbol{H}^{Dis} \end{aligned} \tag{5}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{d_h \times d_h}$ are learnable parameters and $\boldsymbol{A}_1$, $\boldsymbol{A}2 \in \mathbb{R}^{N\times N}$ are temporary matrices projecting from $\boldsymbol{H}^{Spk}$ to $\boldsymbol{H}^{Dis}$ and $\boldsymbol{H}^{Dis}$ to $\boldsymbol{H}^{Spk}$, respectively. Here, $\boldsymbol{H}^{Dis'} \in \mathbb{R}^{N\times d_h}$ can be regarded as a projection from $\boldsymbol{H}^{Spk}$ to $\boldsymbol{H}^{Dis}$, and $\boldsymbol{H}^{Spk'} \in \mathbb{R}^{N\times d_h}$ follows identical principle.

The Whole Process To iteratively refine and exchange discourse structural information and speaker-aware contextual information across multiple consecutive layers, we generalize the calculation process of the initial layer. The detailed procedures are as follows:

$$\begin{aligned} \boldsymbol{H}^{Dis,[l]} &= \textbf{DisGAT}(\boldsymbol{D}^{[l]}, \boldsymbol{E}^{Dis}) \\ \boldsymbol{H}^{Spk,[l]} &= \textbf{SpkGAT}(\boldsymbol{S}^{[l]}, \boldsymbol{E}^{Spk}) \\ \boldsymbol{H}^{Dis',[l]}, \boldsymbol{H}^{Spk',[l]} &= \textbf{Inter}(\boldsymbol{H}^{Dis,[l]}, \boldsymbol{H}^{Spk,[l]}) \\ \boldsymbol{D}^{[l+1]}, \boldsymbol{S}^{[l+1]} &= \boldsymbol{H}^{Dis',[l]}, \boldsymbol{H}^{Spk',[l]} \end{aligned} \tag{6}$$

where $\boldsymbol{D}^{[0]} = \boldsymbol{S}^{[0]} = \boldsymbol{H}^u$ and $l \in [0, L-1]$.

3.3 Emotion Prediction

We obtain the final representation for u_i by concatenating the output $(\boldsymbol{H}^{Dis',[L]}, \boldsymbol{H}^{Spk',[L]})$ of the L-layer DualGATs. The final representation is classified via a Fully-Connected Network (FCN):

$$\begin{aligned} \boldsymbol{l}_i &= \mathit{ReLU}(\boldsymbol{W}_h[\boldsymbol{h}_i^{Dis',[L]} \| \boldsymbol{h}_i^{Spk',[L]}] + \boldsymbol{b}_h) \\ \boldsymbol{p}_i &= \mathit{softmax}(\boldsymbol{W}_l\boldsymbol{l}_i + b_l) \\ \hat{y} &= \mathit{argmax}_{k\in\mathcal{Y}}\boldsymbol{p}_i[k] \end{aligned} \tag{7}$$

where $\hat{y}_i$ is the predicted emotion label for utterance u_i, $\boldsymbol{h}_i^{Dis',[L]}$, $\boldsymbol{h}_i^{Spk',[L]} \in \mathbb{R}^{d_h}$ denote the i^{th} representation in $\boldsymbol{H}^{Dis',[L]}$ and $\boldsymbol{H}^{Spk',[L]}$, $\boldsymbol{W}_h \in \mathbb{R}^{d_h\times 2d_h}$, $\boldsymbol{W}_l \in \mathbb{R}^{d_e\times d_h}$, $\boldsymbol{b}_h \in \mathbb{R}^{d_h}$ and $b_l \in \mathbb{R}^{d_e}$ are learnable parameters of FCN, and d_e denotes the number of emotion labels in the dataset.

3.4 Loss Function

Our training goal is to minimize the following total objective function:

$$\ell_{total} = \ell_{erc} + \lambda\ell_{reg} \tag{8}$$

Table 1: The statistics of four ERC datasets.

Dataset	# Conversations			# Uterrances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5810		1623
MELD	1038	114	280	9989	1109	2610
EmoryNLP	659	89	79	7551	954	984
DailyDialog	11118	1000	1000	87170	8069	7740

where λ is a regularization coefficient. ℓ_{erc} is a standard cross-entropy loss, formulated as:

$$\ell_{erc} = -\sum_{\beta=1}^{B}\sum_{i=1}^{N(\beta)} \log \boldsymbol{p}_{\beta,i}[\boldsymbol{y}_{\beta,i}] \tag{9}$$

where B is the number of conversations, $N(\beta)$ is the number of utterances in the β-th conversation, and $\boldsymbol{y}_{\beta,i}$ is the ground truth label in one-hot form.

4 Experimental Settings

4.1 Datasets

We evaluate our DualGATs on the following four ERC datasets. The statistics of these four datasets are drawn in Table 1.

IEMOCAP (Busso et al., 2008): Each conversation comes from the performance based on the script by two actors. There are 6 emotion labels including *happiness*, *sadness*, *anger*, *frustration*, *excited*, and *neutral*. Since IEMOCAP has no validation set, we follow (Shen et al., 2021b) to use the last 20 conversations in training set for validation.

MELD (Poria et al., 2019a): Scripts collected from the *Friends* TV series. There are 7 emotion labels including *neutral*, *joy*, *surprise*, *sadness*, *anger*, *disgust*, and *fear*.

EmoryNLP (Zahiri and Choi, 2018): Scripts collected from the *Friends* TV series as well. Unlike MELD, its emotion labels include *sad*, *mad*, *scared*, *powerful*, *peaceful*, *joyful*, and *neutral*.

DailyDialog (Li et al., 2017): Daily communications written by human. Its emotion labels are the same as the ones used in MELD.

4.2 Significance Test and Evaluation Metrics

To test the significance of the performance improvement, we conduct a paired t-test with a default level of 0.05 (Koehn, 2004). Following previous methods (Ghosal et al., 2019; Shen et al., 2021a,b), we adopt micro-averaged F1 score excluding the majority class (neutral) for DailyDialog and weighted-average F1 score for the other datasets.

Table 2: Detailed hyper-parameters on each dataset.

Dataset	lr	dropout	batch size	layers
IEMOCAP	1e-4	0.2	16	2
MELD	1e-4	0.3	32	2
EmoryNLP	1e-4	0.1	32	2
DailyDialog	5e-5	0.4	64	3

4.3 Compared Baselines

For a comprehensive performance evaluation, we compare our DualGATs with the following state-of-the-art baselines:

BC-LSTM (Poria et al., 2017), ICON (Hazarika et al., 2018a), DialogueRNN (Majumder et al., 2019), DialogueCRN (Hu et al., 2021), KET (Zhong et al., 2019), DialogueGCN (Ghosal et al., 2019), RGAT (Ishiwatari et al., 2020), DialogXL (Shen et al., 2021a), DAG-ERC (Shen et al., 2021b), CoG-BART (Li et al., 2022), CoMPM (Lee and Lee, 2021), COSMIC (Ghosal et al., 2020), TODKAT[3] (Zhu et al., 2021), DialogueRole (Ong et al., 2022), CauAIN (Zhao et al., 2022), and DisGCN (Sun et al., 2021).

For a fair comparison, baseline+RoBERTa means to use RoBERTa Large (Liu et al., 2019) as an utterance feature extractor as we do. Note that most other baselines natively use pre-trained models as utterance feature extractors, such as DAG-ERC, CoMPM, COSMIC, DialogueRole, and CauAIN use RoBERTa Large, DialogXL uses XLNet (Yang et al., 2019), CoG-BART uses BART (Lewis et al., 2020), and DisGCN uses BERT (Kenton and Toutanova, 2019).

4.4 Implementation Details

Our DualGATs are trained with the Adam optimizer (Kingma and Ba, 2015). We conduct a hyper-parameter search for DualGATs on each dataset according to the F1 score of the validation set. The hyper-parameters to search include learning rate (lr), dropout rate, batch size, and the DualGATs Layer Number (layers) within the ranges of {1e-5,5e-5,1e-4,5e-4,1e-3,5e-3}, {0.0,0.1,0.2,0.3,0.4,0.5}, {8,16,32,64,128}, and {1,2,3,4,5,6}. The details of hyper-parameters for DualGATs on each dataset are shown in Table 2. For other hyper-parameters, the dimension of the feature representation d_u from the RoBERTa is 1024, the dimension of the hidden representation d_h is 300, and the regularization coefficient λ is 0.3. Each training and testing process is run on an NVIDIA A100 GPU with 40GB of memory. Each training process contains 60 epochs, costing at most 50 seconds per epoch. The model with the highest F1 score on the validation set is used to evaluate the test set. The reported results for all our runs are based on the average performance of 5 random runs on the test set.

[3]The sklearn was misused, causing the unusual high performance of MELD and EmoryNLP in TODKAT paper. Thus, we adopt the updated performance from the official Github repository: `https://github.com/something678/TodKat`.

Table 3: The overall performance of all the compared baselines and our DualGATs on four ERC datasets. Bold font denotes the best performance. The marker $*$ refers to significant test $p\text{-}value < 0.05$ comparing with CoMPM, the marker † refers to significant test $p\text{-}value < 0.05$ comparing with CoG-BART, and the marker ‡ refers to significant test $p\text{-}value < 0.05$ comparing with DialogueRole. Moreover, we refer to the results from (Ong et al., 2022) with the marker ♣, from (Shen et al., 2021b) with the marker ♠, from (Bao et al., 2022) with the marker ♢, and the results for the remaining baselines are from original papers.

Models	IEMOCAP	MELD	EmoryNLP	DailyDialog
BC-LSTM♣	54.95	56.87	-	50.24
ICON♣	58.54	-	-	-
DialogueRNN♠	62.75	57.03	-	-
+RoBERTa♠	64.76	63.61	37.44	57.32
DialogueCRN	66.20	58.39	-	-
+RoBERTa♢	66.46	63.42	38.91	-
KET	59.56	58.18	34.39	53.37
DialogueGCN	64.18	58.10	-	-
+RoBERTa♠	64.91	63.02	38.10	57.52
RGAT	65.22	60.91	34.42	54.31
+RoBERTa♠	66.36	62.80	37.89	59.02
DialogXL	65.94	62.41	34.73	54.93
DAG-ERC	68.03	63.65	39.02	59.33
CoG-BART	66.18	64.81	39.04	56.29
CoMPM	66.33	66.52	37.37	60.34
COSMIC	65.28	65.21	38.11	58.48
TODKAT[3]	61.33	65.47	38.69	58.47
DialogueRole	**68.23**	65.34	-	60.95
CauAIN	67.61	65.46	-	58.21
DisGCN	64.10	64.22	36.38	-
DualGATs (Ours)	67.68	**66.90***	**40.69**†	**61.84**‡

5 Results and Discussions

5.1 Main Results

The overall performance of all the compared baselines and our DualGATs on the four datasets is reported in Table 3.

Table 3 shows that when equipped with RoBERTa as a feature extractor, baselines such as DialogueRNN, DialogueCRN, DialogueGCN, and RGAT see considerable improvements. When feature extractors are all based on pre-trained models, graph-based methods such as DialogueGCN

Table 4: Experimental results of ablation study.

Models	IEMOCAP	MELD	EmoryNLP	DailyDialog
DisGAT	64.56	64.23	37.65	58.96
SpkGAT	66.32	64.66	38.34	59.91
DualGATs w/o regularizer	66.70	65.73	39.53	60.93
DualGATs w/o cross attention	66.43	65.46	39.68	60.26
DualGATs (Ours)	**67.68**	**66.90**	**40.69**	**61.84**

or RGAT+RoBERTa, DialogXL, DAG-ERC, CoG-BART, and CoMPM, overall outperform sequence-based methods such as DialogueRNN or Dialogue-CRN+RoBERTa across the four datasets. It indicates that sequence-based methods can not encode the context as effectively as graph-based methods, especially for long-distance contexts. Moreover, when incorporating additional information into the sequence-based or graph-based methods, such as commonsense knowledge in COMSIC, topic information in TODKAT, utterance role in Dialogue-Role, and emotion cause in CauAIN, we see further improvements in overall performance. It indicates that the additional information improves the model's understanding of implicit emotions.

However, these methods neglect the importance of explicitly modeling discourse structure. Compared to focusing on speaker-aware context modeling only, our DualGATs explicitly incorporate discourse structural information by the DisGAT module, so it can capture salient contextual cues that straightforwardly influence emotion. Moreover, GNNs directly only over the parsed discourse dependency graph result in poor performance, such as DisGCN. In contrast, our DualGATs model discourse structure and speaker-aware context simultaneously, achieving competitive performance on the IEMOCAP dataset and reaching a new state of the art on the MELD, EmoryNLP, and DailyDialog datasets compared to all baselines. These results show that our DualGATs effectively integrate discourse structural and speaker-aware contextual information and consider their complementarity for more precise ERC.

5.2 Ablation Study

In this section, we perform ablation studies to analyze the effects of critical modules in our DualGATs, shown in Table 4.

DisGAT only models discourse structure for ERC, which does not work well on datasets that are not sensitive to discourse dependencies due to the complexity of human-human interaction. SpkGAT only models speaker-aware context for ERC and achieves better performance than DisGAT, indicating that for ERC, modeling speaker-aware context is more important than discourse structure. Our DualGATs model both discourse structure and speaker-aware context for ERC and outperform DisGAT and SpkGAT, showing that our DualGATs can simultaneously consider the complementarity of both for more accurate ERC. DualGATs w/o regularizer means we remove the differential regularizer in the interaction module. The results show that the differential regularizer induces the DualGATs to learn more accurate complementary information. DualGATs w/o cross attention denote that we remove the mutual cross-attention transformation in the interaction module so that the DisGAT and SpkGAT modules can not interact. At this point, we concatenate the output representations of the two modules in the last layer to perform emotion prediction. Therefore, the performance drops significantly on four benchmark datasets. Overall, our DualGATs with all modules achieve the best performance.

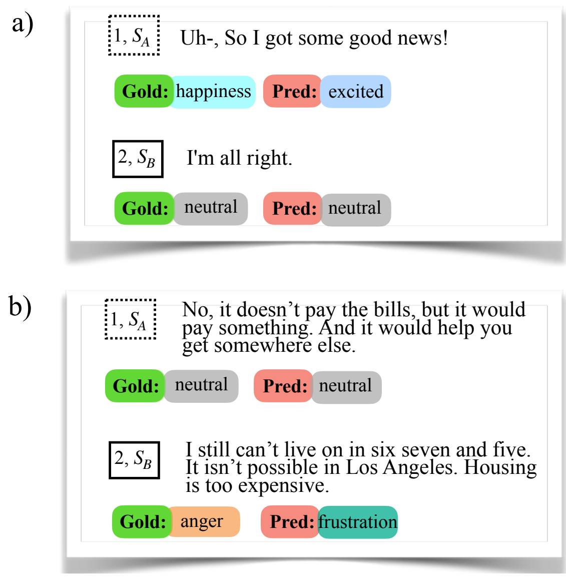

Figure 3: Two real cases of misclassification between *happiness* versus *excited* and *anger* versus *frustration* in the IEMOCAP dataset (Busso et al., 2008).

5.3 Error Analysis

After going through the predicted labels on the four datasets, we find that the following two aspects primarily cause the errors.

Firstly, our DualGATs tend to misclassify utterances of other emotions to *neutral*. This is because most utterances contain *neutral* emotion in the ERC datasets, especially MELD, EmoryNLP, and DailyDialog datasets where the proportion of neutral utterance is 46.95%, 29.95%, and 83.10%,

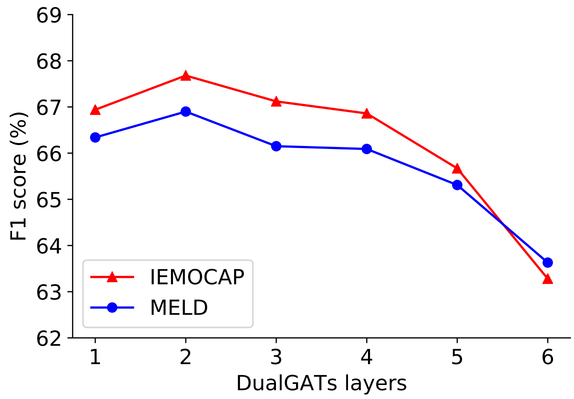

Figure 4: Impact of the number of DualGATs layers.

respectively. These datasets' highly imbalanced class distributions lead to confusion between a few non-neutral utterances and much more neutral ones, restraining the emotion recognition performance.

Secondly, our DualGATs fail to distinguish between emotion pairs that are closely related, such as *happiness* versus *excited*, *anger* versus *frustration*, and *peaceful* versus *joyful*. As shown in Figure 3, we present two cases of misclassification between *happiness* versus *excited* and *anger* versus *frustration* in the IEMOCAP dataset. Taking Figure 3 (a) as an example, it isn't easy to distinguish whether Speaker A was *happiness* or *excited* when she or he said she or he got good news. This misclassification phenomenon between similar emotions has also been reported by (Ghosal et al., 2019; Shen et al., 2021b; Ong et al., 2022).

5.4 Impact of the DualGATs Layer Number

To study the impact of the DualGATs layer number, we evaluate our DualGATs with one to six layers on the IEMOCAP and MELD datasets. As demonstrated in Figure 4, our model with two DualGATs layers performs best. On the one hand, discourse structural information and speaker-aware contextual information might not be refined and exchanged well when the number of layers is small. On the other hand, if there are too many layers, the performance will drop significantly, due to the generation of redundant or compatible representations, canceling important information.

6 Conclusion

In this paper, we propose DualGATs with DisGAT, SpkGAT, and Interaction modules to simultaneously consider the discourse structure's and speaker-aware context's complementarity for more accurate ERC. The DisGAT and SpkGAT incorporate discourse structural and speaker-aware contextual information in parallel. The subsequent interaction module integrates the DisGAT and SpkGAT and effectively exchanges relevant information between the two modules via mutual cross-attention. Experimental results show that our DualGATs outperform previous state-of-the-art baselines on most tested datasets, and further analysis validates the effectiveness of critical modules in DualGATs.

In the future, we will explore the following aspects: (1) Apply our method to similar tasks that need to incorporate the discourse structural and speaker-aware contextual information; (2) Enhance the ability of our method to handle class imbalance or similar emotion problems, such as the introduction of data augmentation or contrastive learning techniques; (3) Deal with domain gap problem when directly using pre-trained deep sequential models to parse conversations in ERC datasets (Dong et al., 2020, 2021).

Limitations

Although our DualGATs simultaneously consider the complementarity of discourse structure and speaker-aware context for more accurate ERC, it requires more computation and a longer training time. The performance of discourse parsing could be more satisfying in the current stage. Moreover, we directly utilize pre-trained deep sequential models to parse dialogues in ERC datasets, which does not address the domain gap problem well.

Ethics Statement

To consider ethical concerns, we describe the following: (1) We conduct all experiments on existing datasets derived from public scientific research. (2) Our work does not involve any sensitive tasks or data. (3) We describe the datasets' statistics and our method's hyper-parameter settings. Our analysis is consistent with the experimental results. (4) We will release our code on GitHub for reproducibility.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727.

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Speaker-Guided Encoder-Decoder Framework for Emotion Recognition in Conversation. *arXiv preprint arXiv:2206.03173*.

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):335–359.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019a. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019b. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.

Feilong Chen, Duzhen Zhang, Minglun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. VLP: A Survey on Vision-language Pre-training. *Int. J. Autom. Comput.*, 20(1):38–56.

Jiaao Chen and Diyi Yang. 2021. Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391.

Sheng Yeh Chen, Chao Chun Hsu, Chuan Chun Kuo, Ting Hao Kenneth Huang, and Lun Wei Ku. 2019. Emotionlines: An emotion corpus of multi-party conversations. In *11th International Conference on Language Resources and Evaluation, LREC 2018*, pages 1597–1601. European Language Resources Association (ELRA).

Kyunghyun Cho, B van Merrienboer, Caglar Gulcehre, F Bougares, H Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Jiahua Dong, Yang Cong, Gan Sun, Zhen Fang, and Zhengming Ding. 2021. Where and How to Transfer: Knowledge Aggregation-Induced Transferability Perception for Unsupervised Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. 2020. What Can Be Transferred: Unsupervised Domain Adaptation for Endoscopic Lesions Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4022–4031.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue Discourse-Aware Graph Model and Data Augmentation for Meeting Summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3808–3814. ijcai.org.

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*, pages 2470–2481.

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)*, pages 154–164.

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*, pages 2594–2604.

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 conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052.

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)*, pages 7360–7370.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn Response Selection using Dialogue Dependency Relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1911–1920.

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)*, pages 397–406.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Joosung Lee and Wooin Lee. 2021. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. *arXiv preprint arXiv:2108.11626*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021a. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1204–1214.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021b. Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis. 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)*, pages 6319–6329.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make BART a good dialogue emotion recognizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11002–11010.

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 Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Ye Liu, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. 2021. Empathetic Dialogue Generation with Pre-trained RoBERTa-GPT2 and External Knowledge. In *Conversational AI for Natural Human-Centric Interaction - 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, volume 943 of *Lecture Notes in Electrical Engineering*, pages 67–81. Springer.

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*.

Zhengyuan Liu and Nancy Chen. 2021. Improving Multi-Party Dialogue Discourse Parsing via Domain Integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127.

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*, volume 33, pages 6818–6825.

Donovan Ong, Jian Su, Bin Chen, Anh Tuan Luu, Ashok Narendranath, Yue Li, Shuqi Sun, Yingzhan Lin, and Haifeng Wang. 2022. Is Discourse Role Important for Emotion Recognition in Conversation?

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.

Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2023. FADO: Feedback-Aware Double COntrolling Network for Emotional Support Conversation. *Knowl. Based Syst.*, 264:110340.

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)*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A Multimodal Multi-Party

Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. 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)*, pages 1551–1560.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2949–2958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

Yunhe Xie, Kailai Yang, Cheng-Jie Sun, Bingquan Liu, and Zhenzhou Ji. 2021. Knowledge-Interactive Network with Sentiment Polarity Intensity-Aware Multi-Task Learning for Emotion Recognition in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2879–2889.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yahan Yu, Bojie Hu, and Yu Li. 2022. GHAN: Graph-Based Hierarchical Aggregation Network for Text-Video Retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5547–5557. Association for Computational Linguistics.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In *IJCAI*, pages 5415–5421.

Duzhen Zhang, Feilong Chen, Jianlong Chang, Xiuyi Chen, and Qi Tian. 2023. Structure Aware Multi-Graph Network for Multi-Modal Emotion Recognition in Conversations. *IEEE Transactions on Multimedia*, 7:1943–1954.

Duzhen Zhang, Xiuyi Chen, Shuang Xu, and Bo Xu. 2020. Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4429–4440.

Duzhen Zhang, Zhen Yang, Fandong Meng, Xiuyi Chen, and Jie Zhou. 2022. TSAM: A Two-Stream Attention Model for Causal Emotion Entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6762–6772.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. CauAIN: Causal Aware Interaction Network for Emotion Recognition in Conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4524–4530. ijcai.org.

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)*, pages 165–176.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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 (Volume 1: Long Papers)*, pages 1571–1582.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section

☒ A2. Did you discuss any potential risks of your work?
Our work does not involve any sensitive data or tasks, and there is no potential risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We will discuss the license at GitHub upon publication.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
4

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.