DialogueEIN: Emotional Interaction Network for Emotion Recognition in Conversations

Anonymous NAACL-HLT 2022 submission

Abstract

Emotion Recognition in Conversations (ERC) is a necessary step for developing empathetic human-computer interaction system. The existing methods on ERC primarily focus on capturing the context-level and speaker-level information from utterances. However, these methods ignore the causes of human emotion change, resulting in insufficient in capturing useful information for emotional prediction. In this work, we propose more explanatory Emotional Interaction Network (DialogueEIN) based on two main stages to capture the contexual information over intra- and inter-speaker dependencies directly from utterances, and to explore and analyze the differentiated contributions over the both kinds of information to boost better understanding of current utterance in conversation. Experimental results on two benchmark datasets demonstrate the effectiveness and superiority of our proposed model.

1 Introduction

Emotion recognition in conversations (ERC) aims at predicting emotion of each utterance in a multi-party conversation. With the growing popularity of human-like artificial intelligence (AI) research, the topic of emotion recognition in conversations has attracted more and more attention from the researchers (Zhang et al., 2019; Li et al., 2020; Zhang and Chai, 2021), especially in recent years. Therefore, accurately identifying the utterance emotion in coversation is a essential step in various fields such as health care (Rashkin et al., 2018; Lin et al., 2019), empathetic chat agents (Althoff et al., 2016; König et al., 2016) and so on.

Unlike vanilla emotion recognition of utterances (Wu et al., 2006; Mohammad and Turney, 2010; Kratzwald et al., 2018), ERC needs to fully consider not only the internal characteristics of utterances, but more importantly, the contextual clues of the utterance in the conversation and the temporality in speakers' turns or speaker-specific information. As a result, ERC is more complex and difficult as natural conversations are usually governed by several different factors or pragmatics (Poria et al., 2019b) that play an important role in a dyadic conversation. Such a scheme is illustrated in Figure 1 that reveals the causes of human emotions in the process of dyadic conversation. We find that these factors, such as the speaker's personality, intention and so on, affect the emotional dynamics of participants through unique interaction. Like most of recent works (Majumder et al., 2019; Ghosal et al., 2019) have been devoted to capturing the context-level and speaker-level cues by deep learning methods. However, these methods do not consider the intrinstic interaction and ignore the flow of contextual and sequential information from utterances in a coversation, resulting in insufficient in understanding of the context.

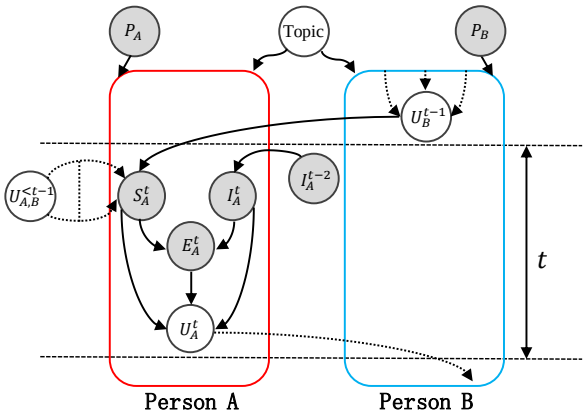


Figure 1: Interaction among different variables during a dyadic conversation between persons A and B. Grey and white circles represent hidden and observed variables, respectively. P represents personality, U represents utterance, S represents interlocutor state, I represents interlocutor intent, E represents emotion.

Further, speaker information is particularly necessary for modelling in the ERC task, because emotional dynamics of conversations consist of two important aspects: intra-speaker (or self-) dependency

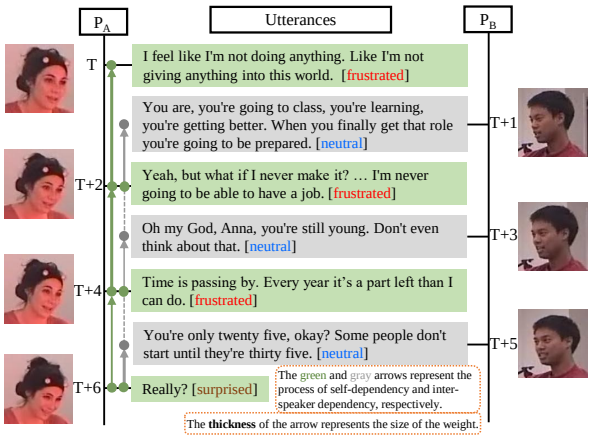


Figure 2: In this conversation, P_A is always frustrated due to the self-dependency before time $T+5$. At time $T+6$, however, P_A is emotionally influenced (i.e., inter-speaker dependency) by P_B at time $T+5$ and thus changes her emotion from *frustrated* to *surprised*.

and inter-speaker dependency (Morris and Keltner, 2000). This phenomenon is illustrated in Figure 2. We can observe that at every turn in the conversation, the individual speaker subconsciously assigns corresponding weights to intra-speaker and inter-speaker dependencies respectively, so as to determine whether the emotion will change. Compared to the recently published works on ERC (Lu et al., 2020; Zhang and Chai, 2021; Li et al., 2022), these methods all ignore this factor.

The *Arguments of Perception and Cognition* (Montemayor and Haladjian, 2017) suggests that our brain's activity architecture is complex, but it can be abstracted into two stages: perceptual activity and cognitive activity, which are independent of each other but affect. In fact, the cognitive process of human beings to objective things is also the fusion process of multiple information (Han, 2010). Motivated by them, in this paper, we propose the more explanatory **E**motional **I**nteraction **N**etwork (**DialogueEIN**) for ERC task. The proposed DialogueEIN consists of two main stages, *i.e*, the interactive representation perception and the interactive representation fusion. In the interactive representation perception stage, we employ three gated recurrent units (GRUs) (Dey and Salem, 2017) to capture the contextual information over intra- and inter-speaker directly from utterances, respectively. All these three different types of GRUs are intertwined to simulate the human-like intrinstic interaction in a recurrent manner. In the interactive representation fusion stage, we first adopt the attention mechanism to retrieve and integrate the emotional clues from the intra- and inter-speaker context, respectively. We surmise that since attention acts on different objects, the two integrated emotional clues will have a certain degree of complementarity. This is confirmed in Section 5.4. Therefore, we employ Transformer (Devlin et al., 2019), which can learn the informative high-dimensional representations from the hidden features, to further analyze the differentiated contribution across the both kinds of information to boost better understanding of current utterance in conversation.

The major contributions are summarized as: 1) A more explanatory DialogueEIN that considering the causes of human emotion change is proposed. 2) The effectiveness of proposed model is demonstrated on two benchmark datasets.

2 Related Work

In 1988, (Minsky, 1988) pointed, "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions". Since then, emotion recognition, as a frontier research, has received increasingly attention from researchers, which can be divided into two phases, i.e., vanilla emotion recognition and emotion recognition in conversations (ERC).

For **vanilla emotion recognition**, some works (Wu et al., 2006; Mohammad and Turney, 2010; Shaheen et al., 2014; Kratzwald et al., 2018) utilized sentiment-lexicon or modern deep learning to extract the internal emotion characteristics of utterances. However, these methods failed to work well in realistic scenario. For **emotion recognition in conversations**, existing works can be further attributed to sequence-based and graph-based methods. The former (Majumder et al., 2019; Wang et al., 2020; Lu et al., 2020; Li et al., 2022) generally employed RNN or Transformer along with attention to capture the context and speaker information over utterances in a conversation. The latter (Ghosal et al., 2019; Lian et al., 2020; Hu et al., 2021) generally adopted graph neural networks (Kipf and Welling, 2016; Schlichtkrull et al., 2018) to capture emotion information by modeling speaker information using dependecies of edges.

Compared to vanilla emotion recognition, ERC has achieved huge performance improvements, yet still faces significant challenges (Poria et al., 2019b). This main reason is that emotions from the human subjective states (Scherer et al., 2001) are highly abstract and require more clues for the

model to understand, unlike the tangible such as animals or objects recognized in the other fields (Zhai et al., 2021). Therefore, this work presents the more explanatory DialogueEIN from the perspective of human emotion derivation.

3 Methodology

3.1 Problem Statement

Figure 2 illustrates one short natural conversation between two people, where each utterance is labeled with an emotional overtones. Formally, given a conversation $U = [u_1, u_2, \ldots, u_N]$ with M participants or speakers $\mathcal{P} = [p_1, p_2, \ldots, p_M]$, where N is the number of utterances in the converstaion and $M \geq 2$. And each utterance u_t is spoken by the speaker $p_{\varnothing(u_t)}$, where $\varnothing$ represents the mapping relationship between the speaker and utterance, and $p_{\varnothing(u_t)} \in \mathcal{P}$. The task of ERC aims to detect the most likely emotion label y_t of the constituent utterance u_t in a conversation U from the emotion category set $\mathcal{Y}$.

3.2 Textual Feature Extraction

Following previous works (Kim, 2014), a multi-channel convolutional neural network (CNN) with filters sizes of {3, 4, 5} and 50 feature maps in each is employed to extract the context-independent n-gram textual features from the transcript of the utterances. Concretely, the 300 dimensional pre-trained 840B GloVe vectors (Pennington et al., 2014) are fed into this networks. Then, a global max-pooling followed by ReLU activation (Nair and Hinton, 2010) further process these feature maps. Finally, these features are concatenated and projected into a d_m dimensional dense layer to form the representation of an utterance. Also, we represent $\{u_t\}_{t=1}^{N}, u_t \in \mathbb{R}^{d_m}$ as the representation of the N utterances.

3.3 Model

Now, we propose our **E**motional **I**nteraction **N**etwork (**DialogueEIN**) for ERC task. The overall framework is illustrated in Figure 3. DialogueEIN is comprised of three main integral components: interactive representation perception (section 3.3.1), interactive representation fusion (section 3.3.2) and emotion classifier (section 3.3.3). The details of the proposed framework are described below.

3.3.1 Interactive Representation Perception

As shown in Figure 1, our daily conversation is governed by interaction among different variables. Some of these variables which can be perceived are observable, while others which can be awared are hidden. We assume that the flow of these variables is limited by two constraints: 1) the intra-speaker dependency; 2) the inter-speaker dependency; where the inter-speaker dependency are directly influenced by the way of interaction among these controlling variables. So, in the perception stage, as shown in Figure 3a, we employ two different types of speaker-GRUs[1] to capture the intra- and inter-speaker dependencies, respectively, and another interaction-GRU to perceive the flow of these variables.

Intra-speaker GRU In the course of a conversation, individual speaker usually has own unique personality, and the speaker's emotion is easily affected by own subjective state (Scherer et al., 2001). So we employ the intra-speaker GRU_P to capture the self-dependency from the adjacent utterances of the same speaker, and expect the GRU_P is aware of the potential personality of individual during the model training.

Based on the current input utterance features $u_t \in \mathbb{R}^{d_m}$, the intra-speaker state $p_{\varnothing(u_t),t-1}$ can be updated to $p_{\varnothing(u_t),t}$ as follows:

$$p_{\varnothing(u_t),t} = GRU_P\left(u_t,\ p_{\varnothing(u_t),t-1}\right), \qquad (1)$$

where D_P is the hidden size of GRU_P cell, $\{p_{\varnothing(u_t),t-1},\ p_{\varnothing(u_t),t}\} \in \mathbb{R}^{D_P}$, and $p_{\varnothing(u_t),t}$ is initialized with null vector for all the participants. Meanwhile, the intra-listener state at the current time t is consistent with that at the previous time $t-1$ as

$$p_{j,t} = p_{j,\,t-1}, \qquad (2)$$

where $j \in [1, \ldots, M]$ and $j \neq \varnothing(u_t)$.

Interaction GRU In the interactive representation perception stage, the interaction GRU is a core step. In this part, we employ the GRU cell GRU_Q to encode those observable variables to adequate understand the contextual information of the utterances in a way that simulates human-like interaction. Intuitively, this modeling method is more interpretable.

Firstly, we use attention mechanism to capture context c_t relevant to the current utterance u_t based on the avaliable representation $q_{*,<t-1}$ of the contextual preceding utterances ($U_*^{<t-1}$) from participants that including both the speaker and the lis-

[1] Taking into account the effectiveness and efficiency, GRU is used here as the basic RNN structure.

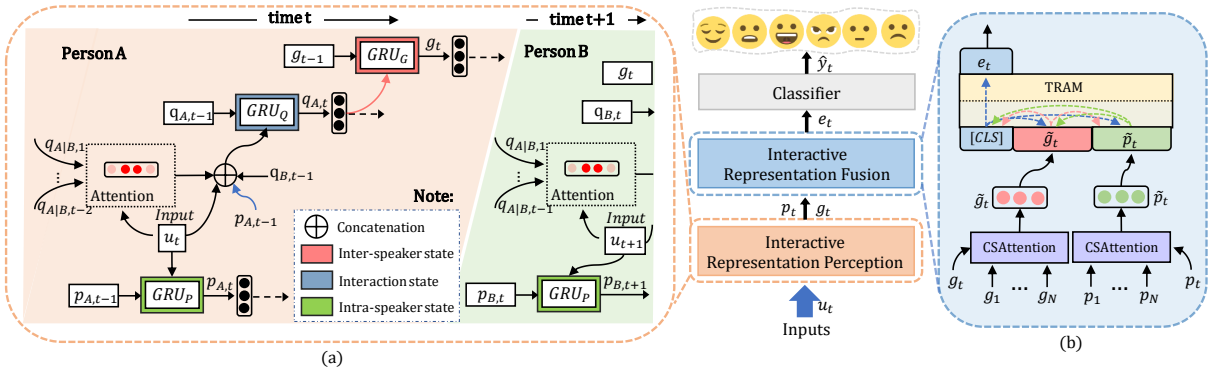


Figure 3: The framework of the DialogueEIN model.

tener. The attention is calculated as:

$$\varphi = \text{softmax}\left(u_t^T W_\varphi [q_{\varnothing(u_1),1}, \ldots, q_{\varnothing(u_{t-2}),t-2}]\right), \quad (3)$$

$$c_t = \varphi [q_{\varnothing(u_1),1}, \ldots, q_{\varnothing(u_{t-2}),t-2}]^T, \quad (4)$$

where $q_{\varnothing(u_i),i}$ is the preceding $i < (t-1)$ hidden state of the interaction GRU_Q, D_Q is the size of the hidden state, $q_{\varnothing(u_i),i} \in \mathbb{R}^{D_Q}$, $c_t \in \mathbb{R}^{D_Q}$, and c_t is assigned a null vector before time $t < 3$. Then, based on the above context representation, the speaker intermediate memory h_t can be obtained by a dense layer as:

$$h_t = W_\tau [u_t \oplus p_{\varnothing(u_t),t-1} \oplus q_{\varnothing(u_{t-1}),t-1} \oplus c_t] + b_\tau, \quad (5)$$

where D_τ is the hidden size of dense layer, $\oplus$ denotes vectors concatenation and $W_\tau \in \mathbb{R}^{D_\tau \times (D_m + D_P + 2D_Q)}$, $b_\tau \in \mathbb{R}^{D_\tau}$ are the learnable parameters. This intermediate memory representation can, on the one hand, effectively fuse different types of information and, on the other hand, reduce the dimension of the interaction state that aims to cut back the memory consumption during computation. Finally, the current interaction state $q_{\varnothing(u_t),t-1}$ can be updated to the new state $q_{\varnothing(u_t),t}$ via the interaction GRU_Q. For speaker,

$$q_{\varnothing(u_t),t} = GRU_Q\left(h_t,\ q_{\varnothing(u_t),t-1}\right), \quad (6)$$

where $q_{\varnothing(u_t),t}$ is initialized with null vector for all the participants. For listener, the update mechanism of interaction state is similar to that of intra-listener state.

Inter-speaker GRU Due to self-dependency, participants in a conversation tend to stick a particular emotional state, unless some external stimulus, usually the other participants, invoke a change (Poria et al., 2019b). In other words, the emotion shift in a conversation has often strong correlation with the inter-speaker dependency. Therefore, the inter-speaker GRU is necessary to perceive the phenomenon.

Now, based on the obtained the interaction state $q_{\varnothing(u_t),t}$ which contains rich context information on all the preceding utterances from the interaction GRU_Q, the inter-speaker state g_t can be updated via the inter-speaker GRU_G as:

$$g_t = GRU_G\left(q_{\varnothing(u_t),t},\ g_{t-1}\right), \quad (7)$$

where D_G is the hidden size of GRU_G, $\{g_{t-1},\ g_t\} \in \mathbb{R}^{D_G}$, and g_t is initialized with null vector, similarly.

Bidirectional Clues Perception Given the current utterance u_t, the above computation (equation (1)-(7)) can be simplied as $p_t, g_t = Perception(u_t; c_t, h_t, q_t)$. In this work, we design the interactive representation perception stage to catpture the context and the speakers information from two directions, *i.e.*, the forward GRUs and backward GRUs. The outputs are represented as:

$$\overrightarrow{p_t}, \overrightarrow{g_t} = \overrightarrow{Perception}\left(u_t; c_t, h_t, \overrightarrow{q_t}\right), \quad (8)$$

$$\overleftarrow{p_t}, \overleftarrow{g_t} = \overleftarrow{Perception}\left(u_t; c_t, h_t, \overleftarrow{q_t}\right), \quad (9)$$

The final representations of the intra-speaker state and the inter-speaker state based on both forward and backward driection at time t are concatenated along the feature dimension, denoted as:

$$p_t = [\overrightarrow{p_t} \oplus \overleftarrow{p_t}], \quad (10)$$

$$g_t = [\overrightarrow{g_t} \oplus \overleftarrow{g_t}], \quad (11)$$

3.3.2 Interactive Representation Fusion

The process of information fusion is essentially a cognitive process of objective things (Han, 2010), which will sublimate the understanding of perceived information to a certain extent. And, the

complementarity of information from different views will be captured by the processing of representation fusion. Therefore, in the interactive representation fusion stage, as shown in Figure 3b, we design the hierarchical module, by cascading the attention and Transformer, to intergrate the emotional clues from intra- and inter-speaker dependencies and to fully explore the internal relationship between these clues to produce a description of the consistency of the predicted current utterance.

Firstly, we employ the **C**ontext **S**ensitive **Attention** (*CSAttention*) to retrieve and integrate the contextual clues from the other surrounding statements in the conversation, due to the inherent problem of poor performance of RNNs in propagating long-term context. For the intra-speaker state:

$$\beta = \text{softmax}\left(p_t^T W[p_1, p_2, \ldots, p_N]\right), \quad (12)$$

$$\widetilde{p_t} = \beta[p_1, p_2, \ldots, p_N]^T, \quad (13)$$

where $W \in \mathbb{R}^{2D_P \times 2D_P}$ is the trainable parameter and $\widetilde{p_t} \in \mathbb{R}^{2D_P}$. We simplify the above computation as:

$$\widetilde{p_t} = \text{CSAttention}(p_t; [p_1, p_2, \ldots, p_N]), \quad (14)$$

Similarly, for the inter-speaker state:

$$\widetilde{g_t} = \text{CSAttention}(g_t; [g_1, g_2, \ldots, g_N]), \quad (15)$$

where $\widetilde{g_t} \in \mathbb{R}^{2D_G}$. $\widetilde{p_t}$ and $\widetilde{g_t}$ hold rich emotional clues over intra- and inter-speaker dependencies, respectively.

Then, in order to uncover the intrinsic relationship between these clues and the extent of their contribution to the correct prediction of the emotion label of the current utterance, we employ the **TRA**nsfor**M**er (*TRAM*) module from BERT (Devlin et al., 2019) to learn the informative high-dimensional representation. In addition, we introduce a special embedding [*CLS*] to make the model free from any bias between them. The input is constituted by adding the [*CLS*] at the head, which is $\{[CLS,\ \widetilde{g_t},\ \widetilde{p_t}]\}$. The process of computation can be denoted as:

$$e_t = \text{TRAM}(CLS,\ \widetilde{g_t},\ \widetilde{p_t}), \quad (16)$$

where $e_t \in 2D_G$ (D_G is equal to D_P for this work) is the fusion vector that indicates the high-dimensional representation from the intra-speaker state and the inter-speaker state, the *TRAM* is comprised of several identical Transformer layers with a final hidden layer named as the BertPooler, and e_t is the output of the last hidden layer at the *CLS* position.

3.3.3 Emotion Classifier

Finally, based on the above output representation e_t from the interactive representation fusion stage, we use an fully-connceted network and a $softmax$ layer as the emotion classifier to predict the emotion label of each utterance, as follows:

$$f_t = \text{GELU}(W_f e_t + b_f), \quad (17)$$

$$\hat{\mathbf{y}}_t = \text{softmax}(W_y f_t + b_y), \quad (18)$$

where $\hat{\mathbf{y}}_t$ is the predicted emotion, $W_f \in \mathbb{R}^{D_f \times 2D_G}$, $b_f \in \mathbb{R}^{D_f}$, $W_y \in \mathbb{R}^{|\mathcal{Y}| \times D_f}$, $b_y \in \mathbb{R}^{|\mathcal{Y}|}$ and $|\mathcal{Y}|$ is the length of the emotion category set $\mathcal{Y}$.

During training, the cross-entropy along with L2-regularization is adpoted as the measure of loss ($\mathcal{L}$). The loss function is defined as:

$$\mathcal{L} = -\frac{1}{\sum_{l=1}^{L} \nu(l)} \sum_{i=1}^{L} \sum_{j=1}^{\nu(i)} \mathbf{y}_{i,j}^l \log\left(\hat{\mathbf{y}}_{i,j}^l\right) + \lambda \|\theta\|_2 \quad (19)$$

where L is the number of samples or conversations, $\nu(i)$ is the number of utterances in the sample i, $\mathbf{y}_{i,j}$ is the one-hot vector for the ground truth, λ is the L2 regularizer weight, and θ is the set of all learnable parameters.

4 Experimental Settings

4.1 Datasets

The proposed DialogueEIN is evaluated on two different benchmark datasets, *i.e.*, **IEMOCAP** (Busso et al., 2008) and **MELD** (Poria et al., 2019a). Detailed statistics for both datasets are reported in the Table 1. In this work, we only focus on the textual modality for emotion recognition in conversations.

IEMOCAP[2] (Busso et al., 2008) The Interactive Emotional dyadic Motion CAPture (IEMOCAP) database is an acted, multi-modal and multi-speaker database that consists of ten unique speakers, belonging to five sessions. The utterances of each conversation are annotated by multiple annotators into six categorical labels, namely $angry$, $happy$, sad, $neutral$, $excited$ and $frustrated$. Following previous works (Majumder et al., 2019), we take the first eight speakers from session one to four and the last two speakers from session five as the training set and the test set, respectively.

MELD[3] (Poria et al., 2019a) Multimodal EmotionLines Dataset (MELD) is a extensions and enhancement of EmotionLines (Hsu et al., 2018) dataset.

[2]https://sail.usc.edu/iemocap/

[3]https://github.com/declare-lab/MELD

Dataset	# Dialogues			# Utterances			# Num. Speakers	# Avg. Length	# Classes
	train	val	test	train	val	test			
IEMOCAP	120		31	5,810		1,623	2	50	6
MELD	1,153	280		11,098	2,610		9↑	10	7

Table 1: The detailed statistics of two datasets; ↑ indicates the maximum number of speakers involved in each conversation.

It contains more than 1400 conversations in which multiple speakers are participated and 13000 utterances from Friends TV series. Every utterance of each conversation is annotated by any of these seven emotions labels, namely $anger$, $disgust$, $sadness$, joy, $neutral$, $surprise$ and $fear$. The pre-defined train/val split provided in the MELD dataset is used in this work.

4.2 Baseline and State-of-the Art methods

For a comprehensive evaluation, we compare the performance of our DialogueEIN with the following baseline methods.

bc-LSTM (Poria et al., 2017) A Bi-directional LSTM is employed to capture the contextual information of utterances from their surroundings in the same conversation.

bc-LSTM+Att (Poria et al., 2017) As the variant of bc-LSTM, the bi-directional contextual LSTM followed by attention mechanism are used to calculate the attention scores by matching the context utterances with the current utterance. Both models are speaker-independent, because they do not consider the speaker's information.

CMN (Hazarika et al., 2018b) CMN adopts the two distinct GRUs followed by attention for two speakers to extract and filter the contextual information from the conversation history. However, this model is difficult to extend to multi-party dataset.

ICON (Hazarika et al., 2018a) ICON, as an extension of CMN, feds the outputs from individual speakers GRUs to another GRU in order to incorporate self and inter-speaker influences in a conversation. Similarly, ICON also is two-party model.

DialogueRNN (Majumder et al., 2019) DialogueRNN employs three GRUs to track the global contextual information, the speaker's state, and the emotional state of each utterance, respectively.

DialogueGCN (Ghosal et al., 2019) DialogueGCN uses GCN (Schlichtkrull et al., 2018) to model the representation of each utterance as node of the graph and the dependence between speakers as edges for emotion recognition in conversations.

BiERU (Li et al., 2022) In this state-of-the art work, a generalized neural tensor block followed by a two-channel classifier is designed to perform context compositionality and emotion classification, respectively. BiERU is also party-ignorant model.

4.3 Implementation Details

The model parameters of the core are set as shown in Table 2. In addition, we choose the Adam as the optimizer with an initial learning rate of {0.0001, 0.0001}, L2 weight decay of {0.00001, 0.0005} and dropout rate of {0.05, 0.2} for *IEMOCAP* and *MELD* datasets, respectively. The batch size is set as 30. For fair comparison with baseline methods, we use the utterance-level textual representation, which is shared by these methods and can be obtained from the open-source project[4].

Dataset	The Perception Stage			The Fusion Stage	
	D_P	D_Q	D_G	layer	head-attention
IEMOCAP	500	500	500	4	8
MELD	150	150	150	3	4

Table 2: The core model parameters setting in both datasets during training.

5 Results and Analysis

5.1 Experimental Results

The experimental results compared with baseline in IEMOCAP and MELD are shown in Table 3. Our proposed DialogueEIN consistently achieves better performance than baseline on both datasets.

For ***IEMOCAP***, our proposed **DialogueEIN** surpasses the best model **BiERU** by 0.4%, 0.9%, and other baseline models by at least 1.7%, 1.5% in terms of accuracy and f1-score, resepctively. Furthermore, our model outperforms in three of the six F1 metrics out of all. In particular, for the *excited*, our model achieves 85.62% accuracy, which is at least 6.7% improvement over all the baseline models. In contrast, **bc-LSTM** and **bc-LSTM+Att** undoubtedly exhibit worst performance because they do not consider speaker information. **CMN**, **ICON**, **DialogueRNN** and **DialogueGCN** model speaker information and global contextual information in different ways. Therefore, they exhibit better performance than the first two baselines. **BiERU** simplifies the step of capturing the context by employing a generalized neural tensor block and a two-channel feature extractor, achieving best performance in all baselines. However, the obtained context also lacks of some emotional clues because the

[4]https://github.com/declare-lab/conv-emotion

Methods	IEMOCAP														MELD	
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
bc-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95	58.93	57.06
bc-LSTM+Att	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	59.54	56.85
CMN	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	**71.13**	60.69	56.56	56.13	-	-
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	58.89	57.06
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	-	58.10
BiERU(SOTA)	**55.44**	31.56	80.19	84.13	**64.73**	59.66	**69.05**	65.25	63.18	74.32	61.06	61.54	66.09	64.59	60.08	56.65
DialogueEIN	25.00	35.64	87.35	**84.58**	55.21	57.92	65.88	**65.69**	**85.62**	**74.96**	64.83	63.17	**66.36**	**65.16**	**60.88**	**58.36**

Table 3: Comparison with the baseline methods on IEMOCAP and MELD dataset using textual modality; Acc. = Accuracy; F1 = Weighted-average F1 score; bold font denotes the best result.

model is speaker-ignorant. Compared with baseline methods, the experimental results demonstrate the effectiveness of our proposed framework.

For ***MELD***, as can be seen in Table 1, the dataset contains conversations, with an average length of about 10 utterances and with up to 9 speakers each conversation, which means that each speaker can be traced back to rarely contextual information. These factors lead to some baseline models showing different performance between MELD and IEMOCAP dataset. From the result in the rightmost part of Table 3, some speaker-ignorant models such as **bc-LSTM**, **BiERU** outperform speaker-dependent model such as **DialogueRNN** in accuracy metric, which demonstrates that it is more difficult to model self-dependency and inter-speaker dependency in MELD. In particular, **DialogueGCN** models speakers information using dependencies of edges in graph networks, inherently improving the contextual understanding of **DialogueRNN** and thus achieves better performance than other baseline methods in F1 metric. Compared with all the baseline methods, our **DialogueEIN** improves by more than 1.3% and 0.4% in terms of accuracy and f1-score, respectively. We surmise that this performance improvement is the result of joint modeling in the perception stage and fusion stage. On the one hand, the interactive representation fusion reduces the difficulty of modeling speaker information, and on the other hand, it also improves the ability to capture context in the perception stage.

5.2 Error Analysis

To further analysis the results, the confusion matrices of classification results in Table 3 is shown in Figure 4a. It can be seen that the *happy* is always misclassified as *excited*. We surmise that this depends on two factors. On the one hand, according to the Valence-Arousal representation (Giannakopoulos et al., 2009), both *happy* and *excited* are positive valence and arousal values emotion, and thus are highly confusing. On the other hand, it is caused by emotional dynamics. In Figure 4b, we illustrate the percentage of successful prediction of several common *emotion-shifts* in the testing set. Observing the histogram carefully, there are significant differences in the predicted *emotion-shift* results between emotion pairs with the different valence and arousal value and those with similar valence and arousal value, e.g., at least 62.5% success from *excited* change to *frustrated* but only 15.15% from *excited* change to *happy*. Further resolution of this issue remains a major challenge in the field of ERC.

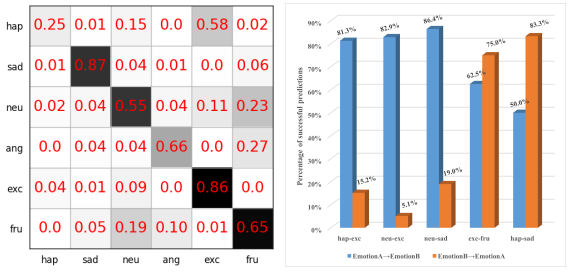
Representation Fusion		Representation Perception		IEMOCAP	
TRAM	*CSAttention*	*Inter-speaker GRU*	*Intra-speaker GRU*	*Acc.*	*F1*
✗	✗	✗	✗	47.13	47.09
✗	✗	✗	✓	57.49	57.42
✗	✗	✓	✗	63.65	63.28
✗	✗	✓	✓	62.6	62.38
✗	✓	✗	✓	59.7	59.6
✗	✓	✓	✗	65.25	63.99
✗	✓	✓	✓	63.4	63.37
✓	✗	✓	✓	61.74	61.34
✓	✓	✓	✓	66.36	65.16

Table 4: The results of ablated DialogueEIN.

5.3 Ablation Study

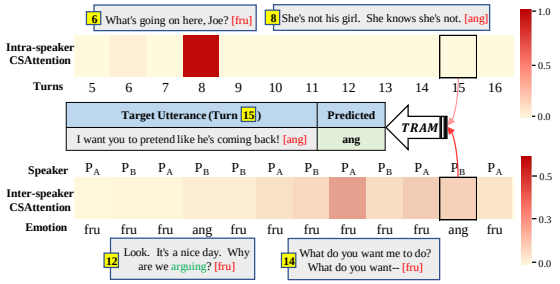
To comprehensively understand the contribution of these two stages, we conduct several ablation studies on IEMOCAP dataset.

As shown in the first block of Table 4, 1) in the first row, when removing perception stage, the performance is dropped sharply (∼24.7%Acc and 24.5%F1). It indicates the necessity for perception stage. 2) In the remaining rows, when only removing either inter- or intra-speaker GRU, the performance is significant decrease and slight increase, respectively. This contrasting results reveal inter-speaker GRU is more important and contains richer contextual cues that trigger emotion than

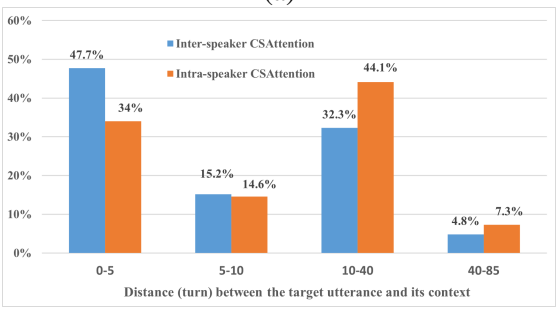


(a) (b)

Figure 4: (a) A set of confusion matrices for DialogueEIN on IEMOCAP. (b) Hiatogram of successfully prediction having a emotion-shift from previous turn on IEMOCAP (e.g., from *EmotionA* change to *EmotionB*).



(a)



(b)

Figure 5: (a) The case study. (b) Highest CSAttention score preference results between test utterance and its context including past and future utterances.

intra-speaker GRU. And, a simple linear layer cannot effectively fuse the two kinds of information.

As shown in the second block, 1) in the first three rows, when adding the CSAttention module, the performance has a certain degree of increase. The phenomenon shows the module can further integrate context from relevant surrounding representations based on attention score. 2) In the last two rows, when only adding the TRAM module, the performance has a slightly decline. In contrast, when TRAM and CSAttention are jointly modeled, there is a significant performance improvement over CSAttention only. In addition, for the *emotion-shifts* of the same speaker, the prediction success probability of the model with TRAM is 4.2% higher than that without TRAM. It indicates that the TRAM, as an indispensable part of the representation fusion stage, can learn informative high-dimensional representation for better understanding current utterance when the obtained features contain sufficient emotional clues.

5.4 Case Study

From Section 5.3 we notice that the fusion stage containing CSAttention and TRAM plays a crucial role, so we have carried out in-depth mining. Figure 5b shows the highest CSAttention score preference distribution over the distance between target utterance and attended utterances. As expected, we observe that the inter-speaker CSAttention ($\sim$ 47.7%) prefers to attend to the local context that are within 5 turns away from themselves. On the contrary, the intra-speaker CSAttention ($\sim$ 44.1%) pays more attention to the long-range context that are 10 to 40 turns away from themselves. This reveals the complementary nature of the contextual cues they capture. Figure 5a shows a complementary case from IEMOCAP dataset. The intra-speaker CSAttention perceives self negative emotion by attending to 8^{th} and 6^{th} turns. Meanwhile, the inter-speaker CSAttention is aware of the *arguing* with P_A by focusing on 12^{th} and 14^{th} turns. Despite the absence of negative expressions throughout 15^{th} turn, after TRAM refining of the integrated clues from intra- and inter-speaker CSAttention, our DialogueEIN still makes the correct *emotion-shift*, i.e., correctly infers the *angry* from previous *frustrated*, where the BiERU misclassified as *frustrated*.

6 Conclusion

This paper attempted to capture and mine emotional clues from two stages of interactive representation perception and interactive representation fusion for emotion recognition in conversations (ERC). We proposed the more explanatory emotional interaction network (DialogueEIN) that first perceived intra- and inter-speaker dependencies directly from the utterance, and then fully mined their intrinsic relationships in order to facilitate better understanding of the current utterance. It achieves comparable performance in two benchmark ERC datasets. Future work will explore the performance of DialogueEIN on multimodal emotion recognition tasks.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Trans. Assoc. Comput. Linguistics*, 4:463–476.

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. *Lang. Resour. Evaluation*, 42(4):335–359.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Rahul Dey and Fathi M. Salem. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *IEEE 60th International Midwest Symposium on Circuits and Systems, MWSCAS 2017, Boston, MA, USA, August 6-9, 2017*, pages 1597–1600. IEEE.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 154–164. Association for Computational Linguistics.

Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. 2009. A dimensional approach to emotion recognition of speech from movies. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, pages 65–68. IEEE.

Richard Langton Gregory. 1970. The intelligent eye.

Chongzhao Han. 2010. *Multi-source information fusion*. Tsinghua University Press.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2594–2604. Association for Computational Linguistics.

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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2122–2132. Association for Computational Linguistics.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

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 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5666–5675. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Alexandra König, Linda E. Francis, Aarti Malhotra, and Jesse Hoey. 2016. Defining affective identities in elderly nursing home residents for the design of an emotionally intelligent cognitive assistant. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2016, Cancun, Mexico, May 16-20, 2016*, pages 206–210. ACM.

Bernhard Kratzwald, Suzana Ilic, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Decision support with text-based emotion recognition: Deep learning for affective computing. *CoRR*, abs/1803.06397.

Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *CoRR*, abs/2003.01478.

Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2022. Bieru: Bidirectional emotional recurrent unit

for conversational sentiment analysis. *Neurocomputing*, 467:73–82.

Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. 2020. Conversational emotion recognition using self-attention mechanisms and graph neural networks. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2347–2351. ISCA.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics.

Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4078–4088. International Committee on Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825. AAAI Press.

Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Carlos Montemayor and Harry H Haladjian. 2017. Perception and cognition are largely independent, but still affect each other in systematic ways: arguments from evolution and the consciousness-attention dissociation. *Frontiers in psychology*, 8:40.

Michael W Morris and Dacher Keltner. 2000. How emotions work: The social functions of emotional expression in negotiations. *Research in organizational behavior*, 22:1–50.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress.

Barbara M Newman and Philip R Newman. 2020. *Theories of adolescent development*. Academic Press.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 873–883. Association for Computational Linguistics.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

K. Scherer, A. Schorr, and T. Johnstone. 2001. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oup Usa.

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, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Shadi Shaheen, Wassim El-Hajj, Hazem M. Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *2014 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2014, Shenzhen, China, December 14, 2014*, pages 383–392. IEEE Computer Society.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 186–195. Association for Computational Linguistics.

Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. 2006. Emotion recognition from text using semantic labels and separable mixture models. *ACM Trans. Asian Lang. Inf. Process.*, 5(2):165–183.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2021. Scaling vision transformers. *CoRR*, abs/2106.04560.

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 *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5415–5421. ijcai.org.

Haidong Zhang and Yekun Chai. 2021. Coin: Conversational interactive networks for emotion recognition in conversation. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 12–18.