# DialogueEIN: Emotion Interaction Network for Dialogue Affective Analysis

**Yuchen Liu**[1], **Jinming Zhao**[2], **Jingwen Hu**[1], **Ruichen Li**[1], **Qin Jin**[1*]
[1] School of Information, Renmin University of China
[2] Qiyuan Lab, Beijing, China
{liuyuchen_alfred, hujingwen_benja, ruichen, qjin}@ruc.edu.cn
zhaojinming@qiyuanlab.com

## Abstract

Emotion Recognition in Conversation (ERC) has attracted increasing attention in the affective computing research field. Previous works have mainly focused on modeling the semantic interactions in the dialogue and implicitly inferring the evolution of the speakers' emotional states. Few works have considered the emotional interactions, which directly reflect the emotional evolution of speakers in the dialogue. According to psychological and behavioral studies, the emotional inertia and emotional stimulus are important factors that affect the speaker's emotional state in conversations. In this work, we propose a novel Dialogue Emotion Interaction Network, **DialogueEIN**, to explicitly model the intra-speaker, inter-speaker, global and local emotional interactions to respectively simulate the emotional inertia, emotional stimulus, global and local emotional evolution in dialogues. Extensive experiments on four ERC benchmark datasets, IEMOCAP, MELD, EmoryNLP and DailyDialog, show that our proposed DialogueEIN considering emotional interaction factors can achieve superior or competitive performance compared to state-of-the-art methods. Our codes and models are released[1].

## 1 Introduction

Emotion Recognition in Conversation (ERC), aiming to recognize the emotional status of each utterance in a conversation, has attracted increasing research attention in recent years. It has rich application potentials in emotional support, mental health, and legal trials etc (Poria et al., 2019).

Unlike traditional emotion recognition based on isolated utterances, conversation context modeling is very important for the ERC task (Poria et al., 2019). Different approaches have been proposed to model the context of a conversation. For example, CMN (Hazarika et al., 2018b) and ICON (Hazarika et al., 2018a) use speaker-specific and global recurrent networks to model the semantic context in a dialogue, and multi-hop memory networks are used to generate the summaries for prediction. DialogueGCN (Ghosal et al.) models the speaker-specific semantic interaction via designing different relations based on graph networks. DialogXL (Shen et al., 2020) utilizes the dialog-aware self-attention mechanism in a transformer structure to capture intra- and inter-speaker dependencies. These works only focus on modeling the dialogue semantic context. Recently, some works have considered the dialogue emotional evolution and proposed several methods to model the emotional context. DialogueRNN (Majumder et al., 2019) and

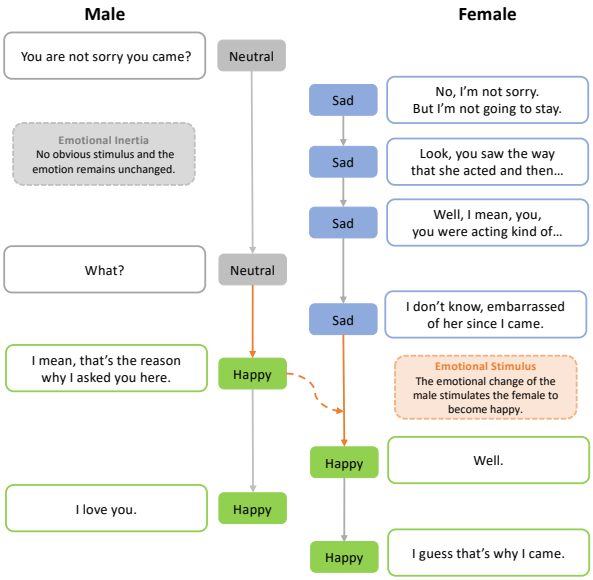


Figure 1: Illustration of emotional interaction in an example dialogue from the IEMOCAP dataset. The grey arrow line indicates that the speaker's emotional state remains unchanged, and the orange arrow line indicates that the speaker is stimulated and the emotion changes, corresponding to emotional inertia and emotional stimulus respectively.

*Corresponding Author
[1]https://github.com/AIM3-RUC/DialogueEIN



*Proceedings of the 29th International Conference on Computational Linguistics*, pages 684–693
October 12–17, 2022.

COSMIC (Ghosal et al., 2020) use separate GRUs to model global and speaker-specific semantic context and utilize another GRU to track the evolution of global emotional states. CESTa (Wang et al., 2020) adopts the Conditional Random Field (CRF) to learn the global emotional consistency in the conversation. IEIN (Lu et al., 2020) proposes to model the global emotion interaction with emotion embeddings and an RNN-based iterative structure. However, these works only consider modeling the global emotional evolution, while ignoring the important speaker-aware emotional dependencies related to emotional inertia and emotional stimulus.

According to psychological and behavioral studies, emotional inertia and emotional stimulus are important factors that affect the speaker's emotional state in dialogues. Emotional inertia (Kuppens et al., 2010; Koval et al., 2015) means that in the absence of sufficient external stimulus, the speaker tends to keep the emotional state unchanged within a dialogue. Emotional stimulus refers to another phenomenon in which a subject's emotion can be aroused and affected by external events which can be words, facial expression, speech intonation or even emotions of the interlocutor (Brosch et al., 2010). Figure 1 illustrates the emotional inertia and emotional stimulus in an example. In this dialogue, the female speaker feels sad at the beginning because of some misunderstanding, but as the misunderstanding is resolved, both speakers appear to be happy. In the first two turns of the dialogue, the emotional states of the male and female speakers remain unchanged (**emotional inertia**). However, after listening to the explanation, the male speaker shows happiness, and his emotional change stimulates the female speaker to become happy as well (**emotional stimulus**). We simply call such emotional inertia and emotional stimulus information as the emotional interaction context in the dialogue, which measures how a person's emotion affects his own or his interlocutor's emotion.

In this paper, we propose a novel Dialogue Emotion Interaction Network(DialogueEIN), to explicitly model the emotional interaction context in conversations for ERC tasks. DialogueEIN mainly consists of a semantic interaction network and an emotional interaction network, where the former aims to capture dialog-level semantic context representations based on a transformer structure, and the latter aims to model the emotional interaction context including intra-speaker emotional inertia, inter-speaker emotional stimulus, global- and local emotional interactions through four corresponding types of dialog-aware self-attention mechanism respectively. We carry out experiments on four ERC benchmark datasets, including IEMOCAP, MELD, EmoryNLP and DailyDialog. The experiment results show that our proposed DialogueEIN achieves state-of-the-art performance on all datasets, which indicates that emotional interactions (such as emotional inertia and emotional stimulus) are important for tracking speakers' emotional evolution in conversations.

The main contributions of this work include:

- We propose the Dialogue Emotion Interaction Network DialogueEIN, to explicitly model the emotional interaction context for ERC tasks.
- We design an Emotional Interaction Network for modeling the self emotional inertia, interlocutor's emotional stimulus, global and local evolution of emotional states.
- Extensive experimental results show that our proposed DialogueEIN achieves the state-of-the-art performance on different benchmark datasets.

## 2 Related Work

**Emotion Recognition in Conversation**

Emotion recognition in conversation has attracted much attention in recent years. There have emerged a number of public emotional dialogue datasets, including IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2018), EmoryNLP (Zahiri and Choi, 2018), DailyDialog (Li et al., 2017) etc.

Different approaches for the ERC task have been proposed as well. Most recent works focus on modeling contextual information in the conversation with different structures. C-LSTM (Poria et al., 2017) uses an simple LSTM-based model to encode the global context in a conversation. CMN (Hazarika et al., 2018b) and ICON (Hazarika et al., 2018a) propose structures based on the gated recurrent unit (GRU) and the memory network to capture both global and speaker-specific context information. DialogueGCN (Ghosal et al.) constructs a graph regarding to both temporal and speaker-aware relationship in the dialogue and model the semantic interactions with a relation-aware graph-based network. DAG-ERC (Shen et al., 2021)

constructs a directed acyclic graph and utilizes a graph-based network to model the information flow in the conversation chronologically, which combine the strengths of conventional graph-based and recurrence-based neural models. DialogXL (Shen et al., 2020) uses an XLNet-based structure, improves memory mechanisms of XLNet and proposes four kinds of dialog-aware attention mechanism to encode corresponding semantic context information in the dialogue.

Above mentioned approaches only focus on modeling the semantic context in the conversation, while some other works consider the emotional context and model the global evolution of emotion states with different methods. DialogueRNN (Majumder et al., 2019) employs several GRUs to track the evolution of different states in the dialogue, including the global emotional state. COSMIC (Ghosal et al., 2020) uses a similar structure to track more kinds of dialog-aware states and introduces external commonsense knowledge to improve the performance. CESTa (Wang et al., 2020) introduces Conditional Random Field (CRF) to learn the emotional consistency in the dialogue. IEIN (Lu et al., 2020) utilizes emotion embeddings and an RNN-based interactive structure to model the global emotion interaction in the conversation.

Our framework is closely related to DialogXL and IEIN, where DialogXL proposes to model speaker-aware context information with attention mechanism, and IEIN proposes to model global emotion interaction based on emotion embeddings. DialogueEIN differs from them from the following two aspects: (1) DialogXL uses speaker-aware attention to model the fine-grained word-level semantic interactions, whereas DialogueEIN focuses on modeling the utterance-level speaker-aware emotional interactions, which explicitly tracks the emotion evolution in the conversation rather than the semantic context. (2) IEIN only considers the global emotional dependencies in the conversation, but ignores the emotional inertia of speakers themselves and the emotional stimulus between interlocutors, whereas DialogueEIN models these two types of speaker-aware emotional interaction explicitly.

**Label Embeddings**

The label embeddings are embedding vectors which are trained to learn the latent knowledge about the label categories in classification tasks. They can be considered as the representations of label categories, where each label embedding vector represents one output label category.

The idea of label embeddings has been widely used in various tasks, including multi-class classification (Bengio et al., 2010), zero-short learning (Larochelle et al., 2008) , text classification (Tang et al., 2015) and sequence labeling (Cui and Zhang, 2019). Cui and Zhang (2019) employs label embeddings and builds a hierarchical label attention network to model the dependencies between output labels for sequence labeling task. Lu et al. (2020) introduces the idea of label embeddings into ERC task to learn the knowledge about emotion labels and model the global emotional interaction. Inspired by these works, we employ label embeddings in our proposed DialogueEIN as well to represent different emotion categories, which is called emotion embeddings.

## 3 Method

### 3.1 Problem Definition

A dialogue can be defined as a sequence of utterances, $\{u_1, u_2, ..., u_N\}$, where $N$ is the total number of utterances. Each utterance $u_j$ contains $n_j$ words, $\{w_1^j, w_2^j, ..., w_{n_j}^j\}$, and uttered by speaker $p(u_j)$, where $p$ is a mapping from utterances to corresponding speakers. Each utterance is labeled with a type of emotion $y_j$, the task is to predict the emotion label of each utterance in a dialogue.

Figure 2 illustrates the overall framework of our proposed DialogueEIN, which consists of four key components: Utterance-level Feature Extraction, Semantic Interaction Network, Emotional Interaction Network and Emotion Classification.

### 3.2 Utterance-level Feature Extraction

We employ a pre-trained RoBERTa (Liu et al., 2019) model to extract utterance-level features. Each utterance $u_j$ is padded with a special token $[CLS]$ and fed into the RoBERTa model:

$$X_j = \text{RoBERTa}([CLS], w_1^j, w_2^j, ..., w_{n_j}^j) \quad (1)$$

where $X_j \in \mathbb{R}^{(n_j+1)\times d_b}$ is the output of the last hidden layer of the RoBERTa model and $d_b$ is the hidden size of the RoBERTa model. We take the hidden state at the $[CLS]$ position of $X_j$ and pass it into a linear layer to get the utterance-level feature representation of $u_j$, which is formulated as follows:

$$x_j = W_u X_{j,0} + b_u \quad (2)$$

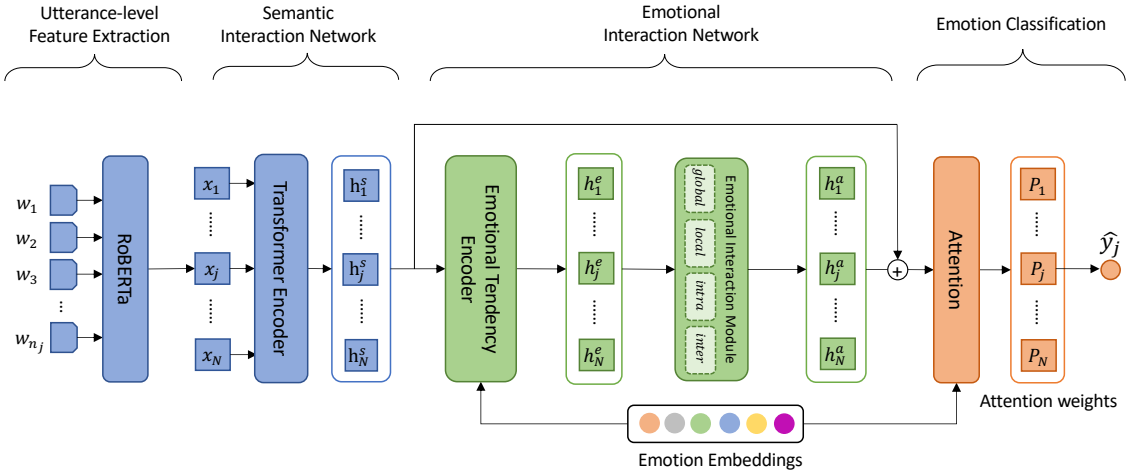


Figure 2: Framework illustration of DialogueEIN, which consists of four key components: Utterance-level Feature Extraction, Semantic Interaction Network, Emotional Interaction Network, Emotion Classification.

where $W_u \in \mathbb{R}^{d_b \times d_u}$, $b_u \in \mathbb{R}^{d_u}$ are learnable parameters and $d_u$ is the dimension of utterance-level feature representations. $x_j$ is the utterance-level feature representation of $u_j$.

### 3.3 Semantic Interaction Network

Since the semantics of each utterance is naturally influenced by other utterances in the dialogue, it is necessary to capture global semantic interaction context of a dialogue. Specifically, in the semantic interaction network, we employ a Transformer (Vaswani et al., 2017) encoder to model semantic interaction in a dialogue.

Given the feature representations of utterances in the dialogue $[x_1, x_2, ..., x_N]$, they are added with a Sinusoidal Position Encoding and then fed into the Transformer encoder. The overall semantic interaction network can be formulated as follows:

$$h^0 = [x_1, x_2, ..., x_N] + \text{PosEnc}(0:N) \quad (3)$$

$$h^s = \text{TRMEncoder}(h^0) \quad (4)$$

where TRMEncoder denotes the transformer encoder model. $h^s$ is the semantic feature representation of the dialogue, which contains the global semantic context information in the dialogue.

### 3.4 Emotional Interaction Network

As mentioned in the introduction, we believe that emotional interaction context, including emotional inertia and emotional stimulus, can benefit the emotion recognition in conversation. We propose an Emotional Interaction Network to model this kind of emotional interaction context, which contains an Emotional Tendency Encoder and an Emotional Interaction Module. Specifically, the Emotional Tendency Encoder can encode the emotional representation of each utterance, which reflects its emotional tendency. Based on these emotional representations, the Emotional Interaction Module models the emotional interactions.

#### 3.4.1 Emotional Tendency Encoder

Inspired by Cui and Zhang (2019) and (Lu et al., 2020), we use emotion embeddings to represent candidate emotion categories and employ a multi-head attention module to capture the emotional tendency of each utterance.

Given the set of candidate emotion labels L = $\{l_1, l_2, ..., l_{|L|}\}$, each label is represented with an embedding:

$$e_i = E^l(l_i) \quad (5)$$

where $E^l$ denotes the emotion embedding lookup table and $e^i$ denotes the embedding of the $i$-th emotion category. As is shown in the structure of Emotional Embeddings in Figure 2, the circles with different colors represent the embeddings of different emotion categories (e.g. happy, sad, anger). These embeddings are initialized randomly and tuned during the model training to learn the latent knowledge about corresponding emotion categories, and can be regarded as the representations of them. The dimension of emotion embeddings is the same as the utterance-level representation, i.e., $e_i \in \mathbb{R}^{d_u}$.

Given the semantic representation $h^s$ and emotion embeddings $e = [e_1, e_2, ..., e_{|L|}]$, a multi-head attention module is applied to them, with $h^s$ as the query and $e$ as the key and the value in the attention mechanism. It is formulated as:

$$h^e = \text{MHA}(h^s, e, e) + h^s \tag{6}$$

Specifically, the multi-head attention module is formulated as:

$$\text{A}(Q, K, V, M) = \text{softmax}(\frac{QK^T}{\sqrt{d_h}} + M)V \tag{7}$$

$$\text{MHA}(Q, K, V, M) = \text{Concat}(head_1, ..., head_n)W^O \tag{8}$$

$$head_i = \text{A}(QW_i^Q, KW_i^K, VW_i^V, M) \tag{9}$$

where $d_h$ denotes the dimension of each attention head, $n$ denotes the number of attention heads, $M$ denotes an attention mask matrix whose elements take value from $\{0, -\infty\}$, and $W_i^Q, W_i^K, W_i^V, W^O$ are trainable parameters. The mask matrix $M$ is set to a null matrix by default.

Since $h^e$ is the result of linear transformations and linear combinations of emotion representations, it contains the explicit emotional information regarding to each utterance and can indicate the emotional tendencies of utterances explicitly. In addition, a residual connection from $h^s$ is added to $h^e$, which means $h^e$ can not only represent the emotional tendency of each utterance, but also carry the semantic information.

### 3.4.2 Emotional Interaction Module

We propose an attention-based module, Emotional Interaction Module, to model the emotional interactions in the conversation. Inspired by (Shen et al., 2020), in order to capture different dependencies and interactions in the conversation, we apply different attention masks to the Emotional Interaction Module.

Specifically, there are four types of attention mechanism are employed, including intra-speaker, inter-speaker, global and local attention, which model the emotional inertia of speakers, the emotional stimulus between interlocutors, the global and local emotional evolution in the dialogue, respectively. The emotional interaction module is formulated as follows:

$$h^{ei} = \text{Concat}\{\text{MHA}_{EI}(h^e, h^e, h^e, m) | m \in M\} \tag{10}$$

$$h^a = \text{LayerNorm}(h^{ei}W^a + b_a + h^s) \tag{11}$$

where $M$=$\{m^{global}, m^{local}, m^{intra}, m^{inter}\}$ denotes global, local, intra-speaker and inter-speaker attention mask respectively, $\text{MHA}_{EI}$ denotes a multi-head attention module, $W^a$ and $b_a$ are trainable parameters.

The formulations and introductions of the four types of self attention masks are shown as follows:

(1) Global Attention Mask: It works the same as the original self-attention mechanism, where each utterance attends to all the utterances in the conversation. Global attention mask is formulated as follows:

$$m_{i,j}^{global} = 0 \tag{12}$$

(2) Local Attention Mask: Each utterance attends to the adjacent utterances within a local window around it. Local attention mask is formulated as follows:

$$m_{i,j}^{local} = \begin{cases} 0, & if\ |i-j| < w/2 \\ -\infty, & otherwise \end{cases} \tag{13}$$

where $w$ is the window size.

(3) Intra-speaker Attention Mask: Each utterance from one speaker only attends to the utterances from the same speaker. Intra-speaker attention mechanism aggregates the emotional state information from each speaker themselves in the dialogue, which models the emotional inertia of them. Intra-speaker attention mask is formulated as follows:

$$m_{i,j}^{intra} = \begin{cases} 0, & if\ p(u_i) = p(u_j) \\ -\infty, & otherwise \end{cases} \tag{14}$$

(4) Inter-speaker Attention Mask: Each utterance from one speaker only attends to the utterances from their interlocutors, contrary to intra-speaker attention. For each speaker's utterance, Inter-speaker attention mechanism aggregates the emotional state information from the interlocutors in the dialogue, which models the emotional stimulus between the speaker and the interlocutors. Inter-speaker attention mask can be formulated as follows:

$$m_{i,j}^{inter} = \begin{cases} 0, & if\ p(u_i) \neq p(u_j)\ or\ i = j \\ -\infty, & otherwise \end{cases} \tag{15}$$

| Dataset | dialogues | | | utterances | | |
|---|---|---|---|---|---|---|
| | train | val | test | train | val | test |
| IEMOCAP | 100 | 20 | 31 | 4830 | 980 | 1623 |
| MELD | 1038 | 114 | 280 | 9989 | 1109 | 2610 |
| EmoryNLP | 713 | 99 | 85 | 9934 | 1344 | 1328 |
| DailyDialog | 11118 | 1000 | 1000 | 87170 | 8069 | 7740 |

Table 1: Data distribution of the four datasets.

### 3.5 Emotion Classification

In order to match the emotion embeddings with corresponding emotion categories, we employ an attention module as the classifier. Given the final representation of utterances $h^e$ and emotion embeddings $e = [e_1, e_2, ..., e_{|L|}]$, the emotion classifier is formulated as follows:

$$P_j = \text{softmax}(e^T W^c h^a) \quad (16)$$
$$\hat{y}_j = \text{argmax}(P_{1:|L|,j}) \quad (17)$$

where $P_j \in \mathbb{R}^{|L| \times N}$ is the attention weights, $W^c$ is the learnable parameter and $\hat{y}_j$ is the prediction of the $j$-th utterance. We regard the attention weights $P_j$ as the predicted probability distribution of emotion labels and directly make predictions based on it. Thus, an utterance with higher attention weight to the emotion embedding vector $e_l$ is more likely to be classified as emotion label $l$, which ensures the matching of emotion embeddings and emotion categories.

Cross-entropy loss is used for model training:

$$L = -\frac{1}{\sum_{k=1}^{T} N(k)} \sum_{i=1}^{T} \sum_{j=1}^{N(i)} \log P_{i,j}[y_{i,j}] \quad (18)$$

where $T$ is the total number of dialogues, $N(i)$ is the number of utterances in dialogue $i$, $y_{i,j}$ and $P_{i,j}$ denote the expected emotion label and the probability distribution of predicted emotion labels of the $j$-th utterance in dialogue $i$ respectively.

## 4 Experiments

### 4.1 Datasets

We carry out evaluations on four ERC benchmark datasets. The distribution of samples in training set, validation set and testing set of these datasets is presented in Table 1. Since IEMOCAP dataset has no validation set, following Shen et al. (2020), we retain the last 20 dialogues in the training set as validation.

**IEMOCAP:** The dataset (Busso et al., 2008) contains 151 two-way conversations from ten speakers in a normal or improvisational way given certain scripts. Each utterance is annotated with an emotion label from six classes, including happy, sad, neutral, angry, excited and frustrated.

**MELD:** The dataset (Poria et al., 2018) contains multi-modal and multi-speaker conversational dialogues from the TV show *Friends*. There are usually three or more speakers in a single conversation. Each utterance is annotated with an emotion label from seven classes, including anger, disgust, fear, joy, neutral, sadness and surprise.

**EmoryNLP:** The dataset (Zahiri and Choi, 2018) is another corpus collected from the TV show *Friends*, which also usually contains more than two speakers in a conversation. Each utterance is annotated with an emotion label from seven classes, including neutral, sad, mad, scared, powerful, peaceful and joyful.

**DailyDialog:** The dataset (Li et al., 2017) collects human-written dyadic conversations from English learning websites with concentrated topics and regulated grammar. Each utterance is annotated with an emotion label from seven classes, including anger, disgust, fear, happiness, sadness, surprise and other.

### 4.2 Implementation Details

We initialize the utterance-level feature extractor with pre-trained RoBERTa models. Specifically, RoBERTa-large model is utilized on MELD, EmoryNLP and DailyDialog datasets. Since the amount of data contained in IEMOCAP dataset is relatively small, we also use a smaller model, RoBERTa-base, on IEMOCAP dataset and only fine-tune the last 4 layers of it during training. We employ an AdamW optimizer (Loshchilov and Hutter, 2018) and a linear learning rate scheduler for model training. The hyper-parameters are tuned on validation set, and two sets of hyper-parameters are used for IEMOCAP and the other three datasets respectively. Specifically, the learning rate of RoBERTa is {2e-5, 5e-6}, the learning rate of the other modules is {1e-4, 5e-5}, the dropout rate is {0.1, 0.1}, the dimension of utterance-level features is {384, 512}, the dimension of feedforward layers is {1024, 2048}, the number of attention heads is {6, 8}, the layers of TransformerEncoder is {4, 4}, the local window size is {15, 5} respectively. The results reported in

the following experiments are based on the average score of 10 random runs.

### 4.3 Comparison with State-of-the-art Methods

We compare our proposed DialogueEIN model to the following state-of-the-art methods. **KET** (Wang et al., 2020) uses a transformer-based structure and external commonsense knowledge to capture the semantic context. **DialogueGCN** (Ghosal et al.) uses graph-based networks to capture conversational dependencies between utterances in dialogues. **DialogueGCN+RoBERTa** means using features based on a more efficient feature extractor RoBERTa instead of GloVe features in DialogueGCN. **DialogXL** (Shen et al., 2020) applies a strong pretrained language model XLNet (Yang et al., 2019) and proposes a dialog-aware self-attention method for modeling the semantic context information. **DAG-ERC** (Shen et al., 2021) constructs a directed acyclic graph and DAGNN (Thost and Chen, 2021) to model the temporal and speaker-aware semantic context. **DialogueRNN** (Majumder et al., 2019) uses several distinct GRUs to model the speaker-specific and global semantic context and another GRU to model the global emotional interaction. It is the first work considering emotional interaction for ERC. **DialogueRNN+RoBERTa** means using features based on a more efficient feature extractor RoBERTa instead of the n-gram features in DialogueRNN. **COSMIC** (Ghosal et al., 2020) proposes a GRU-based structure and uses external commonsense knowledge to capture the semantic context and another GRU to model the global emotional interaction. **CESTa** (Wang et al., 2020) proposes a Transformer- and LSTM-based structure to capture the semantic context and leverages conditional random field (CRF) to model global emotional interaction in conversations.

Table 2 presents the experimental results on the IEMOCAP, MELD, EmoryNLP and DailyDialog four benchmark datasets. DialogueEIN significantly outperforms all other state-of-the-art methods and achieves a new state-of-the-art performance on the IEMOCAP and MELD datasets, which demonstrates its effectiveness of modeling the semantic and emotional context in the dialogue. Please note that the methods in the second block of Table 2 consider global emotional interaction as well. DialogueEIN clearly outperforming these methods indicates that our proposed emotional interaction network with four different emotional interactions can better capture the emotional context. DialogueEIN achieves competitive performance with DAG-ERC on EmoryNLP, which may relate to the fact that dialogues in EmoryNLP are short (3 to 5 utterances on average) and have fewer variations in emotional states, therefore, it is simpler to model by traditional semantic modeling. Additionally, DialogueEIN performs slightly worse than CESTa on DailyDialog, mainly because the dialogues in DailyDialog are short and there are more than 80% of "neutral" emotional states in the dataset.

| | IEMOCAP | MELD | EmoryNLP | DailyDialog |
|---|---|---|---|---|
| | Avg(w) | Avg(w) | Avg(w) | Avg(micro) |
| KET | 59.56 | 58.18 | 33.95 | 53.37 |
| DialogueGCN | 64.18 | 58.10 | - | - |
| +RoBERTa | 64.91 | 63.02 | 38.10 | 57.52 |
| DialogXL | 65.94 | 62.41 | 34.73 | 54.93 |
| DAG-ERC | 68.03 | 63.65 | **39.02** | 59.33 |
| DialogueRNN | 62.75 | 57.03 | - | - |
| +RoBERTa | 64.76 | 63.61 | 37.44 | 57.32 |
| COSMIC | 65.25 | 65.21 | 38.11 | 58.48 |
| CESTa | 67.10 | 58.36 | - | **63.12** |
| DialogueEIN(**Ours**) | **68.93** | **65.37** | 38.92 | 62.58 |

Table 2: ERC performance of different models on four datasets. Micro average F1-score (Avg(micro)) is used on DailyDialog, with the neutral labels excluded. Weighted average F1-score (Avg(w)) is used on other three datasets.

| | IEMOCAP | MELD |
|---|---|---|
| DialogueEIN | **68.93** | **65.37** |
| - intra-speaker attention | 68.63 (0.30↓) | 64.89 (0.48↓) |
| - inter-speaker attention | 68.43 (0.40↓) | 65.10 (0.27↓) |
| - intra-&inter-speaker attention | 67.36 (1.57↓) | 64.84 (0.51↓) |
| - global&local attention | 67.71 (1.22↓) | 64.70 (0.67↓) |
| - Emotional Tendency Encoder | 67.68 (1.25↓) | 64.85 (0.52↓) |
| - Emotional Interaction Network | 66.04 (**2.89**↓) | 64.59 (**0.78**↓) |

Table 3: Ablation Study of DialogueEIN on IEMOCAP and MELD datasets.

### 4.4 Ablation of DialogueEIN

We conduct experiments to ablate the contributions of different components in Emotional Interaction Network, including global, local, intra-speaker, inter-speaker emotional interactions and Emotional Tendency Encoder. The results are shown in Table 3. We can observe that the performance declines obviously when removing each or part of these emotional interactions, which shows that the four emotional interactions are beneficial to ERC. In addition, when the Emotional Tendency Encoder is removed, Emotional Interaction Network would lose the ability to model the emotional context in the conversation, and result in modeling the inter-

| | | IEMOCAP | MELD |
|---|---|---|---|
| 1 | RoBERTa | 63.38 | 62.88 |
| 2 | +TRM | 66.04 | 64.59 |
| 3 | +TRM&Attentions | 67.55 | 64.95 |
| 4 | +TRM&CRF | 67.11 | 64.62 |
| 5 | +TRM&Attentions&CRF | 67.76 | 64.69 |
| 6 | DialogueEIN | **68.93** | **65.37** |

Table 4: Comparasion with RoBERTa-based baselines on IEMOCAP and MELD datasets. We adjust the number of transformer layers so that these models have about the same number of parameters with DialogueEIN.

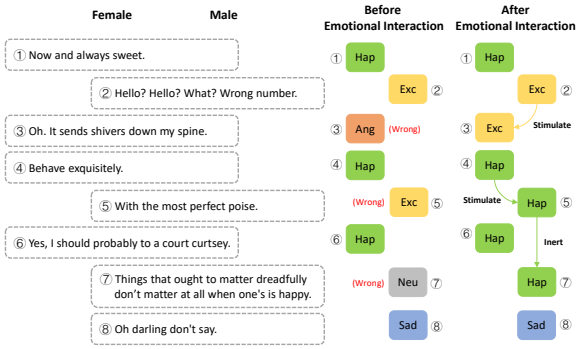


Figure 3: A case study based on IEMOCAP dataset.

actions only based on semantic context. The decline in performance proves that the modeling of emotional context plays an important role in DialogueEIN. Especially, removing the Emotional Interaction Network leads to the worst performance in Table 3. Additionally, the results show that the influence of the Emotional Interaction Network on IEMOCAP and MELD is different (2.9 $\downarrow$ vs 0.78 $\downarrow$), which may be related to the long context (50 utterances per dialogue on average) and the complex emotional evolution of the dialogue in IEMOCAP. It indicates that our proposed Emotional Interaction Network can model long and complex dialogues better.

### 4.5 Comparison with RoBERTa-based Baselines

We adopt a pre-trained language model, RoBERTa, as the utterance-level feature extractor in DialogueEIN. In order to prove that the improvement of DialogueEIN does not come from the increase in the number of parameters and the enhancement of the feature extractor, we propose some RoBERTa-based baselines for comparison: (1) **RoBERTa**: concatenating utterances and feeding them into RoBERTa. (2) **RoBERTa+TRM**: using RoBERTa to extract utterance representations, and feeding them into a transformer encoder. (3) **RoBERTa+TRM+Attentions**: applying dialog-aware attention masks to the transformer encoder. (4) **RoBERTa+TRM+CRF**: following CESTa, adding a CRF layer after the classifier to model the emotional consistency in the dialogue. (5) **RoBERTa+TRM+Attentions+CRF**: further adding dialog-aware attentions into baseline (4).

The results are shown in Table 4. DialogueEIN outperforms all the above mentioned baselines. Comparing row 6 to row 1, the large improvement of DialogueEIN proves that it exploits the full potential of RoBERTa. Comparing row 6 to row 2 and 3, with comparable number of parameters, DialogueEIN outperforms the simple hierarchical transformer structure and this structure with dialog-aware attentions which can model the speaker-specific semantic context. It demonstrates the importance of emotional interaction context modeling in DialogueEIN. Comparing row 6 to row 4 and 5, the results show that CRF can improve the performance to a certain degree in general, especially on IEMOCAP dataset. However, DialogueEIN still outperforms these models, which prove that Emotional Interaction Network is better than CRF in modeling emotional context.

### 4.6 Case Study

Figure 3 shows a conversation from IEMOCAP dataset, and the emotional interaction modeled in DialogueEIN. We extract the attention scores of Emotional Tendency Encoder on these utterances, and use the emotion with maximal score to represent the emotional tendency of each utterance. We also provide the final predictions of these utterances, which are correct predictions as the ground truth. They can represent the emotion prediction of these utterances in DialogueEIN before and after emotional interaction, respectively. As shown in Figure 3, the emotion tendencies before emotional interaction of the $3^{rd}$, $5^{th}$ and $7^{th}$ utterances are different from the final prediction. We illustrate the process that DialogueEIN corrects these errors according to emotional interactions.

(1) The literal meaning of the $3^{rd}$ utterance can be considered as *angry* or *excited*, and the model regards is as *angry* before emotional interaction. However, considering the emotional stimulus from the $2^{nd}$ utterance which is identified as *excited*, it is more rational to identify it as *excited* instead of *angry*.

(2) The $5^{th}$ utterance contains a certain positive emotion literally, but it's hard to distinguish between *happy* or *excited*. However, the female speaker changes to *happy* at the $4^{th}$ utterance, and it stimulates the male speaker's emotion to be *happy* at the $5^{th}$ utterance.

(3) The $7^{th}$ utterance has no obvious emotional tendency literally, and it is regarded as *neutral* before emotional interaction. But when the emotion of the $5^{th}$ utterance is correctly recognized as *happy* and there is no external emotional stimulus from the $6^{th}$ utterance, the $7^{th}$ utterance is finally identified as *happy* according to the emotional inertia.

The above case indicates that DialogueEIN can make more accurate emotion prediction by modeling emotional interaction.

## 5 Conclusion

In this work, we propose a novel emotional interaction Network (DialogueEIN) for Emotion Recognition in Conversation (ERC). DialogueEIN explicitly models emotional inertia, emotional stimulus in a conversation, which most previous works have neglected. DialogueEIN can capture the emotion tendencies of each utterance and model the emotional dependencies based on them. An attention-based Emotional Interaction Network is proposed to measure the emotional interactions, and four types of dialog-aware attentions are employed to simulate emotional inertia, emotional stimulus, global and local evolution of emotional states in the dialogue respectively. Extensive experiments are carried out on IEMOCAP, MELD, EmoryNLP and DailyDialog benchmark datasets. DialogueEIN significantly outperforms other state-of-the-art models on IEMOCAP and MELD datasets, and achieves competitive performance on all four datasets, which proves the effectiveness of the proposed model. Moreover, several ablation studies further explore the structure of DialogueEIN and interpret the use of emotional interaction, which also suggests possible future research directions, such as fusing multimodality to capture emotional stimulus information more accurately, etc.

## 6 Acknowledgments

This work was partially supported by the National Key R&D Program of China (No. 2020AAA0108600), the National Natural Science Foundation of China (No. 62072462).

## References

Samy Bengio, Jason Weston, and David Grangier. 2010. Label embedding trees for large multi-class tasks.

Tobias Brosch, Gilles Pourtois, and David Sander. 2010. The perception and categorisation of emotional stimuli: A review. *Cognition and emotion*, 24(3):377–400.

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.

Leyang Cui and Yue Zhang. 2019. Hierarchically-refined label attention network for sequence labeling. *arXiv preprint arXiv:1908.08676*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation.

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.

Peter Koval, Annette Brose, Madeline L Pe, Marlies Houben, Yasemin Erbas, Dominique Champagne, and Peter Kuppens. 2015. Emotional inertia and external events: The roles of exposure, reactivity, and recovery. *Emotion*, 15(5):625.

Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. 2010. Emotional inertia and psychological maladjustment. *Psychological science*, 21(7):984–991.

Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. 2008. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

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*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

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*, pages 4078–4088.

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.

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. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

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

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2020. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *arXiv preprint arXiv:2012.08695*.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1165–1174.

Veronika Thost and Jie Chen. 2021. Directed acyclic graph neural networks. *arXiv preprint arXiv:2101.07965*.

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.

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*, pages 186–195.

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.

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*.