Knowing What and Why: Causal Emotion Entailment for Emotion Recognition in Conversation

Dazhi Jiang\textsuperscript{a}, Runguo Wei\textsuperscript{a}, Geng Tu\textsuperscript{a}, Hao Liu\textsuperscript{a}, Erik Cambria\textsuperscript{b,∗}

\textsuperscript{a}Department of Computer Science, Shantou University, People’s Republic of China
\textsuperscript{b}School of Computer Science and Engineering, Nanyang Technological University, Singapore

Abstract

Acquisition of clues to motivate emotion is worthy of consideration in emotion recognition in conversation (ERC). In an ideal dialogue system, in fact, it is not only necessary to understand emotions but also to recognize the causes of emotions. Unfortunately, existing research efforts in ERC have ignored the combination of causal emotion entailment for a long time. To this end, we propose an emotion-cause hybrid framework, leveraging causal emotion entailment (CEE) to support the ERC task. Specifically, we integrate the information of the cause clause extracted through the CEE module that triggers emotions into the utterance representations obtained by the ERC model. Moreover, considering the ERC model generally focus on the information near the cause clauses and ignore the context of other clauses during combination, we propose the memory fusion network (MFN) module to supplement the lack of context information. Experimental results show that our framework outperforms state-of-the-art methods on different datasets.

Keywords: Emotion cause, Causal emotion entailment, Memory fusion network, Emotion recognition.

∗Corresponding author
Email addresses: dzjiang@stu.edu.cn (Dazhi Jiang), 20rgwei@stu.edu.cn (Runguo Wei), 19gtu@stu.edu.cn (Geng Tu), 20hliu2@stu.edu.cn (Hao Liu), cambria@ntu.edu.sg (Erik Cambria)
1. Introduction

Emotion recognition in conversation (ERC) plays an important role in AI research, especially for applications related to human-computer interaction such as empathetic dialogue systems but also open-domain and task-oriented chatbots.

Figure 1: Conversation sample where emotions depend on causal emotion entailment. The dotted line indicates that the emotion of the specified utterance is influenced by the cause clause associated with it. The first utterance of Speaker A indicates that he has already been impatient with Speaker B. However, Speaker B apologizes and misinterprets Speaker A’s question. This makes Speaker A annoyed again. Because the emotion-cause pairs between utterances like Speaker A’s first utterance, and Speaker B’s second utterance directly triggered Speaker A’s anger.

The emotion of utterance will be influenced by many factors, such as the conversation context and the causal emotion entailment (CEE). Recent researches on ERC mainly use recurrent neural networks to obtain the dependencies between utterances or use graph-based structure to gain the long-term information. However, these methods neglect to uncover the causes of emotion generation and its utilization, failing to understand and utilize the information associated with emotions entirely. For example, in Figure we display the connection between the cause clauses and emotion clauses. Besides, partial context information will probably be ignored in extracting contextual utterance representation.
To this end, we introduce causal emotion entailment (CEE) and memory fusion network (MFN) into the ERC model. Specifically, in CEE, the utterance is first encoded through the encoding layer and then modeled the inter-clause document through the 2D Window Transformer [13]. After that, the pre-trained 2D Window Transformer model is migrated to the ERC model. Concurrently, the ERC model probably loses information matching the emotion-cause pair in the process of extracting utterance representations. Therefore, we construct the MFN module to fuse the outputs of some intermediate layers of the ERC model to achieve the storage of the probable missing contextual information. Based on this, we combine the framework with the ERC model to observe whether the results are improved and explore the rationality of each module through ablation experiments. The experimental results show that the proposed framework achieves state-of-the-art results on the benchmark dataset. Additionally, generalization analysis, ablation study, case study, and error analysis prove the effectiveness and appropriateness of the CEE and MFN in our framework. The main contributions of this paper are as follows:

- For the first time we combine the CEE module with ERC task, so that the model can use the information of the cause clause associated with the utterance used for emotion prediction.

- We propose the MFN module to solve the problem that utterances probably lose information related to the cause clause during the encoding process. The experimental results show the scientific and robustness of the network.

- We conduct extensive experiments on two different benchmark datasets. The experimental results show that our framework achieves state-of-the-art on the benchmark dataset.

The remaining sections of this paper are organized as follows: Section 2 presents recent relevant work; Section 3 illustrates the ERC task definition; Section 4 presents the proposed methodology; Section 5 lists experimental settings; Section 6 discusses experimental validation on the effectiveness of the proposed approach; finally, Section 7 offers concluding remarks.
2. Related works

**ERC:** (14) proposed that emotion analysis is an interdisciplinary field, which involves psychology, cognitive science and deep learning. Considering the dynamic interaction between speakers, (10, 7) used a recurrent neural network to model different speakers to obtain context information. Due to recurrent neural networks’ long-term information propagation problem, DialogueGCN (11) and DAG-ERC (12) employed graph convolution neural network and directed acyclic graph to model the dialogue context and simulate the information interaction between speakers, respectively. To enrich the utterance representation, KET (15) and COSMIC (16) introduced external knowledge into emotion analysis model by using Knowledge Graph, such as ConceptNet (17), SenticNet (18), and COMET (19), while TODKAT (20) carried out topic detection, and integrated commonsense into transformer to obtain richer context representation. In addition, the pre-trained model has produced quite good results in some applications of natural language processing. DialogXL (21) introduced XLNet (22) into the emotion analysis model and uses the memory units to store the historical information of the dialogue. However, they cannot deal with the problems of difficulty in distinguishing similar emotions and emotion transfer. Therefore, (23) constructed a hybrid learning architecture to alleviate the problems of emotion transfer and confusion labeling in conversational emotion. All of the above approaches do not address the emotion classification problem from the perspective of emotion-cause.

**CEE:** To explore the causes of emotion expression, (24) proposed the emotion-cause extraction (ECE) task and formalized it as a word-level sequence annotation problem. Based on this task setting, many machine learning methods (25) and deep learning methods (26) are used to resolve ECE problems. Further, (27) proposed the emotion-cause pair extraction task (ECPE) to extract potential emotion-cause pairs in documents, and formulated a two-step solution. At the same time, (8) constructed the task of recognizing emotion-cause in conversations with an accompanying dataset named RECCON. It includes two different sub-tasks: causal span extraction (CSE) and causal emotion entailment (CEE).

In previous works, such as (28, 29), they used ECE to solve text-based emotion
classification from the perspective of finding emotion-cause and achieved excellent results. To the best of our knowledge, however, no research work combines CEE with ERC.

3. Task Definition

Let $U = \{u_1, u_2, \cdots, u_N\}$ be a conversation, where $N$ is the number of utterances. And there are $M$ speakers $S = \{S_1, S_2, \cdots, S_M\}$. Each utterance $u_i$ is spoken by the speaker $S_{\phi(u_i)}$, where $\phi$ maps the index of the utterance into that of the corresponding its speaker. We also represent $u_i \in \mathbb{R}^{D_m}$ as the feature representation of the utterance. The task of ERC is to predict the emotion labels (happy, sad, neutral, angry, excited, frustrated, disgust, and fear) of each constituent utterances $u_1, u_2, \cdots, u_N$ from the pre-defined emotion labels. The CEE aims to extract all potential pairs of emotions and corresponding causes from the unannotated emotion document in the conversational context. Given a document $d = [u_1, u_2, \cdots, u_i, \cdots, u_d]$, the goal of CEE task is to extract a set of emotion-cause pairs: $P = \{(u^e, u^c), \cdots\}$ where $u^e$ is an emotion clause and is the corresponding cause clause.

4. Methodology

4.1. The Overall Framework

In this section, we present the overall ERC framework. The framework consists of the CEE module, the ERC model, and the MFN module. An overall structure of the proposed framework is shown in Figure 2.

4.2. Causal Emotion Entailment

On account of CEE can correlate emotion-cause with contextual conversation, we apply the CEE module to ERC task. Specifically, the 2D Window Transformer is used in the pre-trained process. The given utterances are divided into several windows according to window size. The 2D Window Transformer model the relationship between clauses to get better clause representation. 2D Window Transformer has $N$ encoder layers. Each layer consists of a window attention and a feed-forward layer. The 2D
Figure 2: The overall framework. The part (a) is the MFN module. For a given conversation, we encode the utterances using the Transformer to obtain \( x_{1,1} \) and feed it into the MFN module with the intermediate state vectors \( x_{1,2}, x_{1,3}, x_{1,4}, x_{1,5} \) obtained from the ERC model. The part (c) is a pre-trained model. The part (b) is the ERC model, and the part of the ERC model that extracts the utterance context representation is used as a whole as the contextual reasoning module. We concatenate \( x_{1,4} \) with \( x_{1,5} \) and feed it into the first MLP. The dotted lines indicate the direction of data propagation from different modules to each other. The solid lines mean the direction of information propagation between different nodes within each module, where the solid lines in part (c) indicate the process of encoding and decoding the words in the utterance.

Window Transformer is utilized as the encoder layer of the CEE module. Each utterance pair \((u_i, u_j)\) is fed into embedding layer to get the representation \( W_{i,j} \). Firstly, \( W_{i,j} \) is calculated by window attention which is multi-head self-attention. The \( W_{i,j} \) is fed into three linear layers to calculate the query vector \( q_{i,j} \), key vector \( k_{i,j} \) and the value vector \( v_{i,j} \).

\[
q_{i,j} = W_{i,j}W_q \quad (1)
\]
\[
k_{i,j} = W_{i,j}W_k \quad (2)
\]
\[
v_{i,j} = W_{i,j}W_v \quad (3)
\]

where \( W_q \in \mathbb{R}^{n \times n}, W_k \in \mathbb{R}^{n \times n} \) and \( W_v \in \mathbb{R}^{n \times n} \) are learned parameters. For the three vector \( q_{i,j}, k_{i,j} \) and \( v_{i,j} \), the weight \( \beta_{i,j} \) and the output of window attention is calculated as follows:

\[
\beta_{i,j} = \text{softmax} \left( \frac{k_{i,j}^T \cdot q_{i,j}}{\sqrt{n}} \right) \quad (4)
\]
\[
z_{i,j} = v_{i,j} \beta_{i,j}^T \quad (5)
\]
where $z_{i,j}$ is the output of window attention. The input for feed-forward layer is $z_{i,j}$ which is fed into a layer that have two identical constructions followed by normalization layer at its output:

$$o_{i,j,1} = \text{dropout} \left( z_{i,j} W_1 + b_1 \right)$$

$$o_{i,j,2} = o_{i,j,1} + \text{dropout} \left( o_{i,j,1} W_2 + b_2 \right)$$

$$o_{i,j} = o_{i,j,2} + \text{norm} \left( o_{i,j,2} \right)$$

where the $\text{norm}$ denotes laynorm layer. $o_{i,j,1}$ and $o_{i,j,2}$ and are the output of the two sublayers, respectively. $o_{i,j}$ is the output of a encoder layer in 2D Window Transformer.

$$W_{i,j}^{l+1} = o_{i,j}^l$$

where the output $W_{i,j}^N$ of the last layer is the representation of utterance pair $(u_i, u_j)$ extracted by 2D Window Transformer. The relative position modeling is used to learn the representation of clauses pair, and ranks the candidate clauses.

By saving the pre-trained weight and transferring this model, we can convert low-level clause representation to high-level representation, which contains information about the cause evoking the clause. The utterance representation extracted by the pre-trained model is expressed as $md_i$.

### 4.3. Memory Fusion Network

The ERC model loses partial information of other clauses except for the cause clause in the process of extracting utterance representation. Thus, we design the MFN
module to capture the context information. The MFN module have \( N \) layers and the structure of a layer is shown in Figure 3(a). Specially, the input of MFN module is \((x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5})\), and there are two pathways for information fusion. In the right to left pathway, the output \( p_{i,k}^l \) of each node is calculated as follows:

\[
p_{i,k}^l = \text{cell} (w_{k,1} \cdot x_{i,k} + w_{k,2} \cdot p_{i,k+1}^l)
\]

(10)

where \( p_{i,5}^l = x_{i,5} \) and \( w_{k,i} \) is trainable weight that can be a scalar. \text{cell} is the node of MFN module. In the right to left pathway, the output of each node is calculated as follows:

\[
p_{i,k}^o = \text{cell} (w'_{k,1} \cdot x_{i,k} + w'_{k,2} \cdot p_{i,k}^l + w'_{k,3} \cdot p_{i,k-1}^o)
\]

(11)

where \( p_{i,1}^o = x_{i,1} \) and \( w'_{k,i} \) is trainable weight that can be a scalar. \( k \) is the index of the cell in the MFN module and the structure of \text{cell} is shown in Figure 3(b). This utterance representation is fed into BiLSTM, which is followed by the norm layer and activation layer. The output of these cells for the input \( x_i \) can be computed as:

\[
c_i = \text{norm}(x_i + \text{BiLSTM}(x_i))
\]

(12)

\[
p_i^o = \text{ReLU}(x_i + c_i)
\]

(13)

where the norm layer is LayerNorm and the activation layer uses \text{ReLU} function. The output of the previous layer is the input of the next layer.

\[
x_{i,k}^{l+1} = p_{i,k}^l
\]

(14)

where \( l \) is the index of the layer. The output of the final layer are concatenated as the output of the whole memory fusion network. In general, given \((x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5})\), the vector \( m_{fi} \) extracted by of MFN module can be defined as:

\[
m_{fi} = MFN(x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i}, x_{5,i})
\]

(15)

where the \( m_{fi} \) is the output of the MFN module.

### 4.4. Emotion Classifier

Based on the output vectors \( m_{di}, m_{fi} \) obtained from the MFN module and the CEE module, respectively. We concatenate them with the vector \( mc_i \) obtained from the ERC
Table 1: The statistics of datasets. Statistics of splits and evaluation metrics used in different datasets. Neutral classes constitutes to 83% of the DailyDialog dataset which is excluded when calculating the Micro F1 score.

<table>
<thead>
<tr>
<th>Datasets</th>
<th># Conversations</th>
<th># Utterances</th>
<th># Classes</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Val</td>
<td>Test</td>
<td>Train</td>
</tr>
<tr>
<td>IEMOCAP</td>
<td>120</td>
<td>12</td>
<td>31</td>
<td>5810</td>
</tr>
<tr>
<td>DailyDialog</td>
<td>11,118</td>
<td>1,000</td>
<td>1,000</td>
<td>87,832</td>
</tr>
</tbody>
</table>

model last layer and fuse them using MLP to gain the utterance representation $o_i$.

$$o_i = MLP ([md_i; mf_i; mc_i])$$ (16)

where the $o_i$ is the final representation which is fed into the emotion classification layer. Then the emotion classification layer is used for emotion prediction:

$$\hat{y}_i = softmax(W_o o_i + b_o)$$ (17)

We use the cross entropy loss function to calculate the loss value of each module:

$$loss = -\frac{1}{\sum_{i=1}^L c(i) \sum_{i=1}^L y_{i,k}^l \log(\hat{y}_{i,k}^l)}$$ (18)

where $L$ is the number of conversation. $c(i)$ is the number of utterance in the conversation $i$. $y_{i,k}^l$ and $\hat{y}_{i,k}^l$ are the true label of utterance $i$ in conversation $l$ and the possibility of predicting the result of category $k$, respectively.

5. Experimental Settings

5.1. Datasets

We evaluate our framework on the following datasets: IEMOCAP [30], DailyDialog [31]. The detailed statistics of the datasets are reported in Table 1.
IEMOCAP (30) is a multimodal dataset for emotion recognition that contains videos of two-way conversations of ten unique speakers. The utterances are annotated with one of six emotion labels, namely happy, sad, neutral, angry, excited, and frustrated.

DailyDialog (31) is a multi-turn dataset that contains the topics of our daily life and the human-written daily communications. There are seven emotion labels in this dataset: angry, disgust, fear, joy, neutral, sadness, surprise.

Because of the uneven distribution of the DailyDialog dataset, the percentage of utterances with the neutral label is 83%, so we use Micro F1 and Macro F1 excluding the neutral samples. We follow the previous research (7) to use average Accuracy (Acc.) and Weighted F1 on IEMOCAP dataset. In this paper, we use MaF represents the Macro F1, WF represents the Weighted F1 and MiF represents the Micro F1.

5.2. Baselines

To evaluate the proposed framework, we compared it with the following baselines. KET (15) introduces external commonsense knowledge into the transformer by using self-attention and graph-attention mechanisms. VHRED (32) uses a pre-trained sentence encoder and simulating the inter-sentence context through transfer learning to identify emotion. DialogueRNN (7) models context and speaker separately using GRU to obtain global context dependencies and speaker dependencies, meanwhile using global GRU for speaker-to-speaker interaction. DialogueGCN (11) models different speakers using GCN pairs after capturing contextual information, separately. And classify the emotions of utterance representations by attention mechanism. COSMIC (16) employs RoBERTa (33) to extract the data of this paper and introduce commonsense knowledge like mental state, causality, etc. Using the pre-trained model COMET (19) and feed them into the emotion analysis model. DialogueCRN (34) processes utterances representation by using BiLSTM and attention mechanism to simulate the human cognitive. SKAIG (35) introduces commonsense knowledge into the graph structure. DAG-ERC (12) combines traditional graph-based models with recursive-based neural models.
5.3. Feature Extraction

We employ the pre-trained model 840B GloVe \cite{6} to obtain the utterance representation with a dimension of 300. The extracted utterance representations are then fed into a network consisting of a convolutional layer, maximum pooling and fully connected layers to extract text features. The final vector with a dimension of 100 is used as the text feature.

In addition to using GloVe \cite{6} as an encoder, the pre-trained model RoBERTa \cite{33} is also applied for extract context-independent text features. The output vector of the final layer in RoBERTa is used as text feature.

5.4. Hyperparameters Settings

We conduct hyperparameters search for our proposed framework on IEMOCAP and DailyDialog datasets. We employ Adam optimization with a batch size of 32, epochs of 20, the learning rate of 1e-4, L2 weight decay of 2e-4 and dropout of \{0.3, 0.2\}. The number of 2D Window Transformers encoder layers is 3 in the CEE module. The number of layers in the MFN module is 2.
Table 2: The experimental results. The results in bold are the best performing ones under each column. The best values are highlighted in bold.

<table>
<thead>
<tr>
<th>Models</th>
<th># IEMOCAP</th>
<th># DailyDialog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc.</td>
<td>WF</td>
</tr>
<tr>
<td>GLOVe-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KET</td>
<td>-</td>
<td>59.56</td>
</tr>
<tr>
<td>VHRED</td>
<td>-</td>
<td>58.60</td>
</tr>
<tr>
<td>Iterative</td>
<td>-</td>
<td>64.37</td>
</tr>
<tr>
<td>DialogueRNN</td>
<td>63.03</td>
<td>62.50</td>
</tr>
<tr>
<td>DialogueGCN</td>
<td>65.25</td>
<td>64.18</td>
</tr>
<tr>
<td>DialogueCRN</td>
<td>65.25</td>
<td>65.21</td>
</tr>
<tr>
<td>ours(GLOVe)</td>
<td>65.93</td>
<td>65.98</td>
</tr>
<tr>
<td>RoBERTa-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa</td>
<td>-</td>
<td>63.38</td>
</tr>
<tr>
<td>COSMIC</td>
<td>-</td>
<td>65.28</td>
</tr>
<tr>
<td>SKAIG</td>
<td>-</td>
<td>66.96</td>
</tr>
<tr>
<td>DAG-ERC</td>
<td>-</td>
<td>68.08</td>
</tr>
<tr>
<td>ours(RoBERTa)</td>
<td>69.01</td>
<td>69.07</td>
</tr>
</tbody>
</table>

6. Results and Discussions

6.1. The Role of Causal Emotion

To prove the validity of CEE in the ERC model, we conduct one of the most classical models BiLSTM for emotion recognition, where the pre-trained model is used to extract the text features. The results are shown in Figure 4. Compared with using BiLSTM only for emotion classification, the experimental results of BiLSTM+CEE are improved.

6.2. Experimental Results

We compare the our framework based on DialogueCRN model with the baselines in Table 2. As expected, our framework outperforms all the baselines.

On the IEMOCAP dataset, we achieve new state-of-the-art Acc. of 69.01 and WF of 69.07. Compared with the previous work, our framework gains 3.76%, 0.99% in terms of Acc. and WF. On the DailyDialog dataset, our framework gets a 1.44%, 0.06% improvement on MaF and MiF. In order to explain the gaps in experimental...
Models & # IEMOCAP & # DailyDialog & \\
& Acc. & WF & MaF & MiF & \\
BiERU & 63.22 & 63.52 & 29.35 & 52.79 & \\
ours+BiERU & 60.57 & 60.45 & 39.30 & 56.37 & \\
DialogueRNN & 64.20 & 64.21 & 39.69 & 56.19 & \\
ours+DialogueRNN & 66.42 & 66.37 & 51.29 & 58.59 & \\
DialogueCRN & 66.54 & 66.11 & 52.25 & 58.28 & \\
ours+DialogueCRN & 69.01 & 69.07 & 53.39 & 59.81 & \\

Table 3: The experimental results of generalization analysis. The best values are highlighted in bold.

Datasets & BiLSTM & BiERU & DialogueRNN & DialogueCRN & \\
& # IEMOCAP & 9.74e-3 & 7.84e-3 & 9.46e-3 & 3.67e-8 & \\
& # DailyDialog & 4.57e-6 & 6.58e-6 & 9.85e-3 & 1.67e-2 & \\

Table 4: The results of significance tests for generalization analysis (P-Value) (Legends: All P-Values are less than 0.05).

6.3. Generalization Analysis

Inspired by the results of CEE+BiLSTM, we extend our framework to the DialogueCRN model. To prove the effectiveness of combining the CEE module and MFN module with the ERC model, two emotion recognition models (BiERU and DialogueRNN) are used to do generalization analysis. The results are shown in Table 3. As can be seen from Table 3, the results of using the RoBERTa to extract feature are better than using GloVe to extract textual features. Therefore, in this section, RoBERTa is used to extract the text features. Furthermore, the results prove that our framework...
is useful.

For the generalization analysis, we perform significance tests on the compared models. The results are shown in Table 4. The results demonstrate that our framework is significantly different from the comparison model.

6.4. Ablation Study

To check the contribution of different modules, we perform several ablation studies in Table 5 where constituent components are removed respectively. And the ablation study is implemented on the ERC model DialogueCRN. When the MFN and CEE modules are removed one after another, the performance drops slightly. The experimental results show the significance of the MFN and CEE modules for ERC.

Analysis of Memory Fusion Network: As shown in Table 5 when the ERC model utilizes GloVe-based to extract text features, the results are improved on both IEMOCAP and DailyDialog datasets. Meanwhile, when Roberta is used to extracting text features, the results are improved on the IEMOCAP dataset, in which the results of the evaluation indicators Acc. and WF are 67.90 and 68.19, increased by 1.36% and

<table>
<thead>
<tr>
<th>MFN</th>
<th>CEE</th>
<th># IEMOCAP</th>
<th># DailyDialog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(\mathcal{R}_{BERT})</td>
<td>(\mathcal{R}_{RoBERTa})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Acc. MaF WF</td>
<td>Acc. MaF WF</td>
</tr>
<tr>
<td>GloVe-based</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓ X</td>
<td>X</td>
<td>65.25 64.48 65.21</td>
<td>65.25 64.48 65.21</td>
</tr>
<tr>
<td>✓ X</td>
<td>✓</td>
<td>65.68 66.05 65.65</td>
<td>65.68 66.12 65.72</td>
</tr>
<tr>
<td>✓ X</td>
<td>✓</td>
<td>65.99 66.53 66.07</td>
<td>65.99 66.53 66.07</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓</td>
<td>65.93 65.85 65.98</td>
<td>66.17 66.02 66.12</td>
</tr>
<tr>
<td>RoBERTa-based</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓ X</td>
<td>X</td>
<td>66.54 65.81 66.11</td>
<td>66.54 65.81 66.11</td>
</tr>
<tr>
<td>✓ X</td>
<td>✓</td>
<td>67.71 66.81 67.43</td>
<td>68.33 66.07 67.97</td>
</tr>
<tr>
<td>✓ X</td>
<td>✓</td>
<td>67.90 67.11 68.19</td>
<td>67.90 67.11 68.19</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓</td>
<td>68.86 67.04 68.58</td>
<td>69.01 68.15 69.07</td>
</tr>
</tbody>
</table>

Table 5: The experimental results of ablation studies on IEMOCAP and Dailydialog datasets (Legends: \(\mathcal{R}\) represents the CEE Pre-trained Model. The best values are highlighted in bold).
2.08%, respectively. Similarly, the results are improved on the DailyDialog dataset, indicating that the MFN module has strong performance.

**Analysis of Causal Emotion Entailment:** The results show in Table 5. When using CEE in the DialogueCRN model only, whether GloVe or Roberta is used to extract text features, the results are improved. When Roberta is applied to extracting text features, the results of evaluation indexes Acc. and WF on IEMOCAP dataset are 68.33 and 67.97, with an increase of 1.79% and 1.86% respectively. The results of MaF and MiF on the DailyDialog dataset are 49.01 and 58.38, which means promotion is weak.

In case of combining MFN with CEE together, the result is better than that using one of them alone on IEMOCAP and DailyDialog datasets. We can draw a conclusion that ablation with respect to both modules at the same time leads to a higher drop in ERC model performance. That shows that the MFN module and the CEE module can complement each other.

### 6.5. Case Study

In this section, we illustrate a case study on a conversation example from the Dailydialog dataset in Figure 5, which shows the role of the cause clause on the emotion of the utterance. The emotion label of utterance 5 and utterance 7 is easily predicted to be neutral, while the actual label is happiness. As is shown by the solid lines in Figure 5, the cause clauses of utterance 5 are utterance 1 and utterance 3. While the dotted lines indicate that the cause clauses of utterance 7 are utterance 3, utterance 5, and utterance 6. When not performing the CEE, utterance 5 and utterance 7 obtain contextual information by using the ERC model, we will incorrectly predict their emotion as neutral after a few rounds of training. In contrast, with the CEE task, the model enhances the effectiveness of the cause clause on the associated utterance emotion. During model training, the contextual representation of utterance 5 will contain more information related to utterance 1 and utterance 3, and the contextual representation of utterance 7 will contain more information related to utterance 3, utterance 5, and utterance 6, which causes utterance 5 and utterance 7 to be correctly predicted as happiness. The illustrated utterances show that CEE highlights the influence of some clauses on the particular utterance.
Hello, Joanna. You are looking very charming in the new dress.

Thanks. Does it suit me?

Yes, it suits you very well. It certainly is unique. I don’t think I’ve seen anything like it before.

I know. That’s why I bought it. I hate wearing the same styles like everybody else is wearing.

And the necklace. It matches your dress marvelously.

It’s very nice of you to say so. I should say you are glamorous yourself, as a matter of fact.

Thank you for saying so.

Figure 5: Case study on the DailyDialog dataset.

6.6. Error Analysis

Despite our framework performs strongly, it is still defective. The analysis of the experimental results revealed that the model proposed in this paper could not effectively distinguish between similar emotion categories, such as excited, happy, neutral, and frustrated. Figure 6 shows the classification results of our experimental results for these four kinds of samples on the IEMOCAP dataset. A similar situation occurs in the DailyDialog dataset. We suspect that utterances with similar emotions have sim-
ilar semantic information in the extracted features. In addition, the experiments are implemented on text data. The existing research shows that multimodal data can provide more vital information for utterances with non-neutral emotions. For example, the corresponding video data will have a disappointed expression for utterances with sad emotions. Utterance with angry emotion will have a higher pitch. Due to the limitations of the CEE task, not all utterances used for emotion analysis can obtain information about their corresponding cause clauses, which limits the performance of our framework.

7. Conclusion

In this paper, we proposed a framework combining CEE and MFN to improve the emotion analysis ability in ERC. In particular, our framework focuses on the cause clause that triggers emotions via the CEE module and helps the issue of ignoring the context of other clauses when the CEE module is combined with the ERC model by the MFN module. Our framework achieves state-of-the-art results on two benchmark datasets. As future work, we plan to dedicate more attention to incorporating multimodal information into this framework while fusing commonsense information effectively.

Acknowledgments

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046). The research work is also funded by the National Natural Science Foundation of China (62106136, 61902231), Natural Science Foundation of Guangdong Province (2019A1515010943), The Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Special Projects in Artificial Intelligence)(2019KZDZX1030), 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (2020LKSFG04D), Science and Technology Major Project of Guangdong Province (STKJ2021005), and the Opening Project of GuangDong Province Key Laboratory of Information Security Technology(2020B1212060078).


[10] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, Icon: Interactive conversational memory network for multimodal emotion detection, in: Pro-


[37] W. Li, W. Shao, S. Ji, E. Cambria, Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis, Neurocomputing 467 (2022) 73–82.