DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations

Dou Hu[1], **Lingwei Wei**[2,3], **Xiaoyong Huai**[1]
[1] National Computer System Engineering Research Institute of China
[2] Institute of Information Engineering, Chinese Academy of Sciences
[3] School of Cyber Security, University of Chinese Academy of Sciences
{hudou18, weilingwei18}@mails.ucas.edu.cn
huaixy@sina.com

Abstract

Emotion Recognition in Conversations (ERC) has gained increasing attention for developing empathetic machines. Recently, many approaches have been devoted to perceiving conversational context by deep learning models. However, these approaches are insufficient in understanding the context due to lacking the ability to extract and integrate emotional clues. In this work, we propose novel Contextual Reasoning Networks (DialogueCRN) to fully understand the conversational context from a cognitive perspective. Inspired by the Cognitive Theory of Emotion, we design multi-turn reasoning modules to extract and integrate emotional clues. The reasoning module iteratively performs an intuitive retrieving process and a conscious reasoning process, which imitates human unique cognitive thinking. Extensive experiments on three public benchmark datasets demonstrate the effectiveness and superiority of the proposed model.

1 Introduction

Emotion recognition in conversation (ERC) aims to detect emotions expressed by the speakers in each utterance of the conversation. The task is an important topic for developing empathetic machines (Zhou et al., 2020) in a variety of areas including social opinion mining (Kumar et al., 2015), intelligent assistant (König et al., 2016), health care (Pujol et al., 2019), and so on.

A conversation often contains contextual clues (Poria et al., 2019) that trigger the current utterance's emotion, such as the cause or situation. Recent context-based works (Poria et al., 2017; Hazarika et al., 2018b; Majumder et al., 2019) on ERC have been devoted to perceiving situation-level or speaker-level context by deep learning models. However, these methods are insufficient in understanding the context that usually contains rich emotional clues. We argue they mainly suffer from the following challenges. 1) **The extraction of emotional clues**. Most approaches (Hazarika et al., 2018a,b; Jiao et al., 2020b) generally retrieve the relevant context from a static memory, which limits the ability to capture richer emotional clues. 2) **The integration of emotional clues**. Many works (Majumder et al., 2019; Ghosal et al., 2019; Lu et al., 2020) usually use the attention mechanism to integrate encoded emotional clues, ignoring their intrinsic semantic order. It would lose logical relationships between clues, making it difficult to capture key factors that trigger emotions.

The *Cognitive Theory of Emotion* (Schachter and Singer, 1962; Scherer et al., 2001) suggests that cognitive factors are potently determined for the formation of emotional states. These cognitive factors can be captured by iteratively performing the intuitive retrieving process and conscious reasoning process in our brains (Evans, 1984, 2003, 2008; Sloman, 1996). Motivated by them, this paper attempts to model both critical processes to reason emotional clues and sufficiently understand the conversational context. By following the mechanism of *working memory* (Baddeley, 1992) in the cognitive phase, we can iteratively perform both cognitive processes to guide the extraction and integration of emotional clues, which imitates human unique cognitive thinking.

In this work, we propose novel Contextual Reasoning Networks (DialogueCRN) to recognize the utterance's emotion by sufficiently understanding the conversational context. The model introduces a cognitive phase to extract and integrate emotional clues from the context retrieved by the perceive phase. Firstly, in the perceptive phase, we leverage Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks to capture situation-level and speaker-level context. Based on the above context, global memories can be obtained to storage different contextual information. Sec-

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 7042–7052
August 1–6, 2021. ©2021 Association for Computational Linguistics

ondly, in the cognitive phase, we design multi-turn reasoning modules to iteratively extract and integrate the emotional clues. The reasoning module performs two processes, *i.e.*, an intuitive retrieving process and a conscious reasoning process. The former utilizes the attention mechanism to match relevant contextual clues by retrieving static global memories, which imitates the intuitive retrieving process. The latter adopts LSTM networks to learn intrinsic logical order and integrate contextual clues by retaining and updating dynamic working memory, which imitates the conscious reasoning process. It is slower but with human-unique rationality (Baddeley, 1992). Finally, according to the above contextual clues at situation-level and speaker-level, an emotion classifier is used to predict the emotion label of the utterance.

To evaluate the performance of the proposed model, we conduct extensive experiments on three public benchmark datasets, *i.e., IEMOCAP*, *SEMAINE* and *MELD* datasets. Results consistently demonstrate that our proposed model significantly outperforms comparison methods. Moreover, understanding emotional clues from a cognitive perspective can boost the performance of emotion recognition.

The main contributions of this work are summarized as follows:

- We propose novel Contextual Reasoning Networks (DialogueCRN) to fully understand the conversational context from a cognitive perspective. To the best of our knowledge, this is the first attempt to explore cognitive factors for emotion recognition in conversations.
- We design multi-turn reasoning modules to extract and integrate emotional clues by iteratively performing the intuitive retrieving process and conscious reasoning process, which imitates human unique cognitive thinking.
- We conduct extensive experiments on three public benchmark datasets. The results consistently demonstrate the effectiveness and superiority of the proposed model[1].

2 Methodology

2.1 Problem Statement

Formally, let $U = [u_1, u_2, ..., u_N]$ be a conversation, where N is the number of utterances. And there are M speakers/parties $p_1, p_2, ..., p_M$ $(M \geq 2)$. Each utterance u_i is spoken by the speaker $p_{\phi(u_i)}$, where ϕ maps the index of the utterance into that of the corresponding speaker. Moreover, for each $\lambda \in [1, M]$, we define U_λ to represent the set of utterances spoken by the speaker p_λ, *i.e.*, $U_\lambda = \{u_i \mid u_i \in U$ and u_i spoken by p_λ, $\forall i \in [1, N]\}$.

The task of emotion recognition in conversations (ERC) aims to predict the emotion label y_i for each utterance u_i from the pre-defined emotions $\mathcal{Y}$.

2.2 Textual Features

Convolutional neural networks (CNNs) (Kim, 2014) are capable of capturing n-grams information from an utterance. Following previous works (Hazarika et al., 2018b; Majumder et al., 2019; Ghosal et al., 2019), we leverage a CNN layer with max-pooling to exact context-free textual features from the transcript of each utterance. Concretely, the input is the 300 dimensional pre-trained 840B GloVe vectors (Pennington et al., 2014). We employ three filters of size 3, 4 and 5 with 50 feature maps each. These feature maps are further processed by max-pooling and ReLU activation (Nair and Hinton, 2010). Then, these activation features are concatenated and finally projected onto a dense layer with dimension $d_u = 100$, whose output forms the representation of an utterance. We denote $\{\mathbf{u}_i\}_{i=1}^N, \mathbf{u}_i \in \mathbb{R}^{d_u}$ as the representation for N utterances.

2.3 Model

Then, we propose Contextual Reasoning Networks (DialogueCRN) for emotion recognition in conversations. DialogueCRN is comprised of three integral components, *i.e.,* the perception phase (Section 2.3.1), the cognition phase (Section 2.3.2), and an emotion classifier (Section 2.3.3). The overall architecture is illustrated in Figure 1.

2.3.1 Perception Phase

In the perceptive phase, based on the input textual features, we first generate the representation of conversational context at situation-level and speaker-level. Then, global memories are obtained to storage different contextual information.

Conversational Context Representation. Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) introduces the gating mechanism into recurrent neural networks to

[1] The source code is available at `https://github.com/zerohd4869/DialogueCRN`

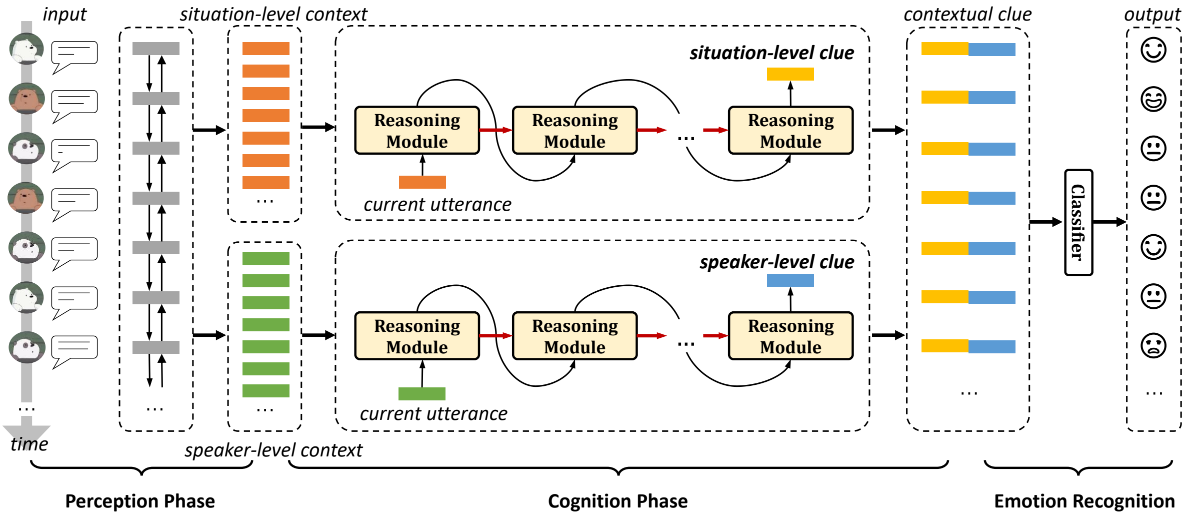


Figure 1: The architecture of the proposed model DialogueCRN.

capture long-term dependencies from the input sequences. In this part, two bi-directional LSTM networks are leveraged to capture situation-level and speaker-level context dependencies, respectively.

For learning the context representation at the situation level, we apply a bi-directional LSTM network to capture sequential dependencies between adjacent utterances in a conversational situation. The input is each utterance's textual features $\mathbf{u}_i \in \mathbb{R}^{d_u}$. The situation-level context representation $\mathbf{c}_i^s \in \mathbb{R}^{2d_u}$ can be computed as:

$$\mathbf{c}_i^s, \mathbf{h}_i^s = \overleftrightarrow{LSTM}^s(\mathbf{u}_i, \mathbf{h}_{i-1}^s), \tag{1}$$

where $\mathbf{h}_i^s \in \mathbb{R}^{d_u}$ is the i-th hidden state of the situation-level LSTM.

For learning the context representation at the speaker level, we also employ another bi-directional LSTM network to capture self-dependencies between adjacent utterances of the same speaker. Given textual features $\mathbf{u}_i$ of each utterance, the speaker-level context representation $\mathbf{c}_i^v \in \mathbb{R}^{2d_u}$ is computed as:

$$\mathbf{c}_i^v, \mathbf{h}_{\lambda,j}^v = \overleftrightarrow{LSTM}^v(\mathbf{u}_i, \mathbf{h}_{\lambda,j-1}^v), j \in [1, |U_\lambda|], \tag{2}$$

where $\lambda = \phi(u_i)$. U_λ refers to all utterances of the speaker p_λ. $\mathbf{h}_{\lambda,j}^v \in \mathbb{R}^{d_u}$ is the j-th hidden state of speaker-level LSTM for the speaker p_λ.

Global Memory Representation. Based on the above conversational context representation, global memories can be obtained to storage different contextual information via a linear layer. That is, global memory representation of situation-level context $\mathbf{G}^s = [\mathbf{g}_1^s, \mathbf{g}_2^s, ..., \mathbf{g}_N^s]$ and that of speaker-level context $\mathbf{G}^v = [\mathbf{g}_1^v, \mathbf{g}_2^v, ..., \mathbf{g}_N^v]$ can be computed as:

$$\mathbf{g}_i^s = \mathbf{W}_g^s \mathbf{c}_i^s + \mathbf{b}_g^s, \tag{3}$$

$$\mathbf{g}_i^v = \mathbf{W}_g^v \mathbf{c}_i^v + \mathbf{b}_g^v, \tag{4}$$

where $\mathbf{W}_g^s, \mathbf{W}_g^v \in \mathbb{R}^{2d_u \times 2d_u}$, $\mathbf{b}_g^s, \mathbf{b}_g^v \in \mathbb{R}^{2d_u}$ are learnable parameters.

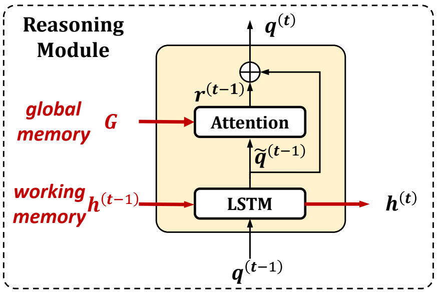


Figure 2: The detailed structure of reasoning module.

2.3.2 Cognition Phase

Inspired by the *Cognitive Theory of Emotion* (Schachter and Singer, 1962; Scherer et al., 2001), cognitive factors are potently determined for the formation of emotional states. Therefore, in the cognitive phase, we design multi-turn reasoning modules to iteratively extract and integrate the emotional clues. The architecture of a reasoning module is depicted in Figure 2.

The reasoning module performs two processes, the intuitive retrieving process, and the conscious reasoning process. In the t-th turn, for the **reasoning process**, we adopt the LSTM network to learn intrinsic logical order and integrate contextual clues in the working memory, which is slower but with human-unique rationality (Baddeley, 1992). That is,

$$\tilde{\mathbf{q}}_i^{(t-1)}, \mathbf{h}_i^{(t)} = \overrightarrow{LSTM}(\mathbf{q}_i^{(t-1)}, \mathbf{h}_i^{(t-1)}), \quad (5)$$

where $\tilde{\mathbf{q}}_i^{(t-1)} \in \mathbb{R}^{2d_u}$ is the output vector. $\mathbf{q}_i^{(t)} \in \mathbb{R}^{4d_u}$ is initialized by the context representation $\mathbf{c}_i$ of the current utterance, *i.e.*, $\mathbf{q}_i^{(0)} = \mathbf{W}_q\mathbf{c}_i + \mathbf{b}_q$, where $\mathbf{W}_q \in \mathbb{R}^{4d_u \times 2d_u}$ and $\mathbf{b}_q \in \mathbb{R}^{4d_u}$ are learnable parameters. $\mathbf{h}_i^{(t)} \in \mathbb{R}^{2d_u}$ refers to the working memory, which can not only storage and update the previous memory $\mathbf{h}_i^{(t-1)}$, but also guide the extraction of clues in the next turn. During sequential flowing of the working memory, we can learn implicit logical order among clues, which resembles the conscious thinking process of humans. $\mathbf{h}_i^{(t)}$ is initialized with zero. t is the index that indicates how many "processing steps" are being carried to compute the final state.

For the **retrieving process**, we utilize an attention mechanism to match relevant contextual clues from the global memory. The detailed calculations are as follows:

$$\mathbf{e}_{ij}^{(t-1)} = f(\mathbf{g}_j, \tilde{\mathbf{q}}_i^{(t-1)}), \quad (6)$$

$$\alpha_{ij}^{(t-1)} = \frac{\exp(\mathbf{e}_{ij}^{(t-1)})}{\sum_{j=1}^{N} \exp(\mathbf{e}_{ij}^{(t-1)})}, \quad (7)$$

$$\mathbf{r}_i^{(t-1)} = \sum_{j=1}^{N} \alpha_{ij}^{(t-1)} \mathbf{g}_j, \quad (8)$$

where f is a function that computes a single scalar from $\mathbf{g}_j$ and $\tilde{\mathbf{q}}_i^{(t-1)}$ (*e.g.*, a dot product).

Then, we concatenate the output of reasoning process $\tilde{\mathbf{q}}_i^{(t-1)}$ with the resulting attention readout $\mathbf{r}_i^{(t-1)}$ to form the next-turn query $\mathbf{q}_i^{(t)}$. That is,

$$\mathbf{q}_i^{(t)} = [\tilde{\mathbf{q}}_i^{(t-1)}; \mathbf{r}_i^{(t-1)}]. \quad (9)$$

The query $\mathbf{q}_i^{(t)}$ will be updated under the guidance of working memory $\mathbf{h}_i^{(t)}$, and more contextual clues can be retrieved from the global memory.

To sum up, given context representation $\mathbf{c}_i$ of the utterance u_i, global memory representation $\mathbf{G}$, and the number of turns T, the whole cognitive phase (Eq.5-9) can be denoted as, $\mathbf{q}_i = Cognition(\mathbf{c}_i, \mathbf{G}; T)$. In this work, we design two individual cognition phases to explore contextual clues at situation-level and speaker-level, respectively. The outputs are defined as:

$$\mathbf{q}_i^s = Cognition^s(\mathbf{c}_i^s, \mathbf{G}^s; T^s), \quad (10)$$

$$\mathbf{q}_i^v = Cognition^v(\mathbf{c}_i^v, \mathbf{G}^v; T^v), \quad (11)$$

where T^s and T^v are the number of turns in situation-level and speaker-level cognitive phases, respectively.

Based on the above output vectors, the final representation $\mathbf{o}$ can be defined as a concatenation of both vectors, *i.e.*,

$$\mathbf{o}_i = [\mathbf{q}_i^s; \mathbf{q}_i^v]. \quad (12)$$

2.3.3 Emotion Classifier

Finally, according to the above contextual clues, an emotion classifier is used to predict the emotion label of the utterance.

$$\hat{\mathbf{y}}_i = softmax(\mathbf{W}_o\mathbf{o}_i + \mathbf{b}_o), \quad (13)$$

where $\mathbf{W}_o \in \mathbb{R}^{8d_u \times |\mathcal{Y}|}$ and $\mathbf{b}_o \in \mathbb{R}^{|\mathcal{Y}|}$ are trainable parameters. $|\mathcal{Y}|$ is the number of emotion labels.

Cross entropy loss is used to train the model. The loss function is defined as:

$$\mathcal{L} = -\frac{1}{\sum_{l=1}^{L} \tau(l)} \sum_{i=1}^{L} \sum_{k=1}^{\tau(i)} \mathbf{y}_{i,k}^l log(\hat{\mathbf{y}}_{i,k}^l), \quad (14)$$

where L is the total number of conversations/samples in the training set. $\tau(i)$ is the number of utterances in the sample i. $\mathbf{y}_{i,k}^l$ and $\hat{\mathbf{y}}_{i,k}^l$ denote the one-hot vector and probability vector for emotion class k of utterance i of sample l, respectively.

3 Experimental Setups

3.1 Datasets

We evaluate our proposed model on following benchmark datasets, *IEMOCAP* (Busso et al., 2008), *SEMAINE* (McKeown et al., 2012), and *MELD* (Poria et al., 2019) datasets. The statistics are reported in Table 1. The above datasets are multimodal datasets with textual, visual, and acoustic features. In this paper, we focus on emotion recognition in textual conversations. Multimodal emotion recognition in conversations is left as future work.

IEMOCAP[2]: The dataset (Busso et al., 2008) contains videos of two-way conversations of ten

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

Dataset	# Dialogues			# Utterances			Avg. Length	# Classes
	train	*val*	*test*	*train*	*val*	*test*		
IEMOCAP	120		31	5,810		1,623	50	6
SEMAINE	63		32	4,368		1,430	72	4*
MELD	1,039	114	280	9,989	1,109	2,610	10	7

* refers to the number of real valued attributes.

Table 1: The statistics of three datasets.

unique speakers, where only the first eight speakers from session one to four belong to the training set. The utterances are annotated with one of six emotion labels, namely *happy*, *sad*, *neutral*, *angry*, *excited*, and *frustrated*. Following previous works (Hazarika et al., 2018a; Ghosal et al., 2019; Jiao et al., 2020b), the validation set is extracted from the randomly shuffled training set with the ratio of 80:20 since no pre-defined train/val split is provided in the *IEMOCAP* dataset.

SEMAINE[3]: The dataset (McKeown et al., 2012) is a video database of human-agent interactions. It is available at AVEC 2012's *fully continuous sub-challenge* (Schuller et al., 2012) that requires predictions of four continuous affective attributes: *Arousal*, *Expectancy*, *Power*, and *Valence*. The gold annotations are available for every 0:2 seconds in each video (Nicolle et al., 2012). Following (Hazarika et al., 2018a; Ghosal et al., 2019), the attributes are averaged over the span of an utterance to obtain utterance-level annotations. We utilize the standard both training and testing splits provided in the sub-challenge.

MELD[4]: Multimodal Emotion Lines Dataset (MELD) (Poria et al., 2019), a extension of the EmotionLines (Hsu et al., 2018), is collected from TV-series Friends containing more than 1400 multi-party conversations and 13000 utterances. Each utterance is annotated with one of seven emotion labels (*i.e.*, *happy/joy*, *anger*, *fear*, *disgust*, *sadness*, *surprise*, and *neutral*). We use the pre-defined train/val split provided in the *MELD* dataset.

3.2 Comparisons Methods

We compare the proposed model against the following baseline methods. **TextCNN** (Kim, 2014) is a convolutional neural network trained on context-independent utterances. **Memnet** (Sukhbaatar et al., 2015) is an end-to-end memory network and update memories in a multi-hop fashion. **bc-LSTM+Att** (Poria et al., 2017) adopts a bidirectional LSTM network to capture the contextual content from the surrounding utterances. Additionally, an attention mechanism is adopted to re-weight features and provide a more informative output. **CMN** (Hazarika et al., 2018b) encodes conversational context from dialogue history by two distinct GRUs for two speakers. **ICON** (Hazarika et al., 2018a) extends CMN by connecting outputs of individual speaker GRUs using another GRU for perceiving inter-speaker modeling. **DialogueRNN** (Majumder et al., 2019) is a recurrent network that consists of two GRUs to track speaker states and context during the conversation. **DialogueGCN** (Ghosal et al., 2019) a graph-based model where nodes represent utterances and edges represent the dependency between the speakers of the utterances.

3.3 Evaluation Metrics

Following previous works (Hazarika et al., 2018a; Jiao et al., 2020b), for *IEMOCAP* and *MELD* datasets, we choose the **accuracy score** (*Acc.*) to measure the overall performance. We also report the **Weighted-average F1 score** (*Weighted-$F1$*) and **Macro-averaged F1 score** (*Macro-$F1$*) to evaluate the model performance on both majority and minority classes, respectively. For the *SEMAINE* dataset, we report **Mean Absolute Error** (*MAE*) for each attribute. The lower *MAE*, the better the detection performance.

3.4 Implementation Details

We use the validation set to tune hyperparameters. In the perceptive phase, we employ two-layer bi-directional LSTM on *IEMOCAP* and *SEMAINE* datasets and single-layer bi-directional LSTM on the *MELD* dataset. In the cognitive phase, single-layer LSTM is used on all datasets. The batch size is set to 32. We adopt Adam (Kingma and Ba, 2015) as the optimizer with an initial learning rate of $\{0.0001, 0.001, 0.001\}$ and L2 weight decay of $\{0.0002, 0.0005, 0.0005\}$ for *IEMOCAP*, *SEMAINE*, *MELD* datasets, respectively. The dropout rate is set to 0.2. We train all models for a maximum of 100 epochs and stop training if the validation loss does not decrease for 20 consecutive epochs.

For results of DialogueGCN and DialogueRNN, we implement them according to the public code[5] provided by Majumder et al. (2019); Ghosal et al. (2019) under the same environment.

[3] https://semaine-db.eu

[4] https://github.com/SenticNet/MELD

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

IEMOCAP			
Methods	*Acc.*	*Weighted-F*1	*Macro-F*1
TextCNN	49.35	49.21	48.13
Memnet	55.70	53.10	55.40
bc-LSTM+Att	56.32	56.19	54.84
CMN	56.56	56.13	54.30
ICON	59.09	58.54	56.52
DialogueRNN	63.03	62.50	60.66
DialogueGCN	64.02	63.65	63.43
DialogueCRN	**66.05**	**66.20**	**66.38**
Improve	**3.2%**	**4.0%**	**4.7%**

Table 2: Experimental results on the *IEMOCAP* dataset.

SEMAINE				
Methods	*MAE*			
	Valence	*Arousal*	*Expectancy*	*Power*
TextCNN	0.545	0.542	0.605	8.71
Memnet	0.202	0.211	0.216	8.97
bc-LSTM+Att	0.189	0.213	0.190	8.67
CMN	0.192	0.213	0.195	8.74
ICON	0.180	0.190	0.180	8.45
DialogueRNN	0.175	0.171	0.181	8.66
DialogueGCN	0.176	0.210	0.193	8.65
DialogueCRN	**0.173**	**0.152**	**0.175**	**8.20**
Improve	**1.1%**	**11.1%**	**2.8%**	**2.9%**

Table 3: Experimental results on the *SEMAINE* dataset.

4 Results and Analysis

4.1 Experimental Results

Table 2, 3 and 4 show the comparison results for emotion recognition in textual conversations. DialogueCRN consistently achieves better performance than the comparison methods on all datasets, while also being statistically significant under the paired t-test (p<0.05).

***IEMOCAP* and *SEMAINE*.** Both *IEMOCAP* and *SEMAINE* datasets have long conversation lengths and the average length is not less than 50. The fact implies that the two datasets contain richer contextual information. **TextCNN** ignoring conversational context obtains the worst performance. **Memnet** and **bc-LSTM+Att** perceive the situation-level context of the current utterance. **CMN** perceives the speaker-level context. Thereby, **Memnet**, **bc-LSTM+Att** and **CMN** slightly outperforms **TextCNN**. **ICON**, **DialogueRNN**, and **DialogueGCN** consider both situation-level and speaker-level context to model the perceptive phase of context. They achieve better performance than the above methods. Compared with baseline methods, **DialogueCRN** can extract and integrate rich emotional clues by exploring cognitive factors. Accordingly, our model obtains more effective performance. That is, as shown in Table 2 and 3, for the *IEMOCAP* dataset, **DialogueCRN** gains 3.2%, 4.0%, 4.7% relative improvements over the previous best baselines in terms of *Acc.*, *Weighted-F*1, and *Macro-F*1, respectively. For the *SEMAINE* dataset, **DialogueCRN** achieves a large margin of 11.1% MAE for the *Arousal* attribute.

MELD			
Methods	*Acc.*	*Weighted-F*1	*Macro-F*1
TextCNN	59.69	56.83	33.80
bc-LSTM+Att	57.50	55.90	34.84
CMN	-	54.50	-
ICON	-	54.60	-
DialogueRNN	59.54	56.39	32.93
DialogueGCN	59.46	56.77	34.05
DialogueCRN	**60.73**	**58.39**	**35.51**
Improve	**2.0%**	**2.9%**	**1.9%**

Table 4: Experimental results on the *MELD* dataset.

***MELD*.** From Table 1, the number of speakers of each conversation in the *MELD* dataset is large (up to 9), and the average length of conversations is 10. The shorter conversation length of the *MELD* dataset indicates it contains less contextual information. From the result in Table 4, interestingly, **TextCNN** ignoring conversational context achieves better results than most baselines. It indicates that it is difficult to learn useful features from perceiving a limited and missing context. Besides, **DialogueGCN** leverages graph structure to perceive the interaction of multiple speakers, which is sufficient to perceive the speaker-level context. Thereby, the performance is slightly improved. Compared with baselines, **DialogueCRN** enables to perform sequential thinking of context and understand emotional clues from a cognitive perspective. Therefore, it achieves the best recognition results, *e.g.*, 2.9% improvements on *Weighted-F*1.

4.2 Ablation Study

To better understand the contribution of different modules in DialogueCRN to the performance, we conduct several ablation studies on both *IEMOCAP* and *SEMAINE* datasets. Different modules that model the situation-level and speaker-level context in both perceptive and cognitive phases are removed separately. The results are shown in Table 5. When cognition and perception modules are removed successively, the performance is greatly

Cognition		*Perception*		**IEMOCAP**			**SEMAINE**			
Situation Context	*Speaker Context*	*Situation Context*	*Speaker Context*	*Acc.*	*Weighted-F*1	*Macro-F*1	*MAE*			
							Valence	*Arousal*	*Expectancy*	*Power*
✓	✓	✓	✓	**66.05**	**66.20**	**66.38**	**0.173**	**0.152**	**0.175**	**8.201**
✓	×	✓	✓	64.26	64.43	62.86	0.173	0.162	0.181	8.253
×	✓	✓	✓	63.28	63.8	63.31	0.174	0.165	0.179	8.201
×	×	✓	✓	63.22	63.37	62.08	0.177	0.171	0.180	8.237
×	×	×	✓	63.50	63.68	62.40	0.192	0.213	0.195	8.740
×	×	✓	×	60.07	60.14	59.58	0.194	0.212	0.201	8.900
×	×	×	×	49.35	49.21	48.13	0.545	0.542	0.605	8.710

Table 5: Experimental results of ablation studies on *IEMOCAP* and *SEMAINE* datasets.

declined. It indicates the importance of both the perception and cognition phases for ERC.

Effect of Cognitive Phase. When only removing cognition phase, as shown in the third block of Table 5, the performance on the *IEMOCAP* dataset decreases 4.3%, 4.3% and 6.5% in terms of *Acc.*, *Weighted-F*1, and *Macro-F*1, respectively. And on the *SEMAINE* dataset, the *MAE* scores of *Valence*, *Arousal*, and *Expectancy* attributes are increased by 2.3%, 12.5% and 2.9%, respectively. These results indicate the efficacy of the cognitive phase, which can reason based on the perceived contextual information consciously and sequentially. Besides, if removing the cognitive phase for either speaker-level or situation-level context, as shown in the second block, the results decreased on both datasets. The fact reflects both situational factors and speaker factors are critical in the cognitive phase.

Effect of Perceptive Phase. As shown in the last row, when removing the perception module, the performance is dropped sharply. The inferior results reveal the necessity of the perceptive phase to unconsciously match relevant context based on the current utterance.

Effect of Different Context. When removing either situation-level or speaker-level context in both cognitive and perceptive phases, respectively, the performance has a certain degree of decline. The phenomenon shows both situation-level and speaker-level context play an effective role in the perceptive and cognitive phases. Besides, the margin of dropped performance is different on both datasets. This suggests speaker-level context plays a greater role in the perception phase while more complex situation-level context works well in the cognitive phase. The explanation is that it is limited to learn informative features from context by intuitive matching perception, but conscious cognitive reasoning can boost better understanding.

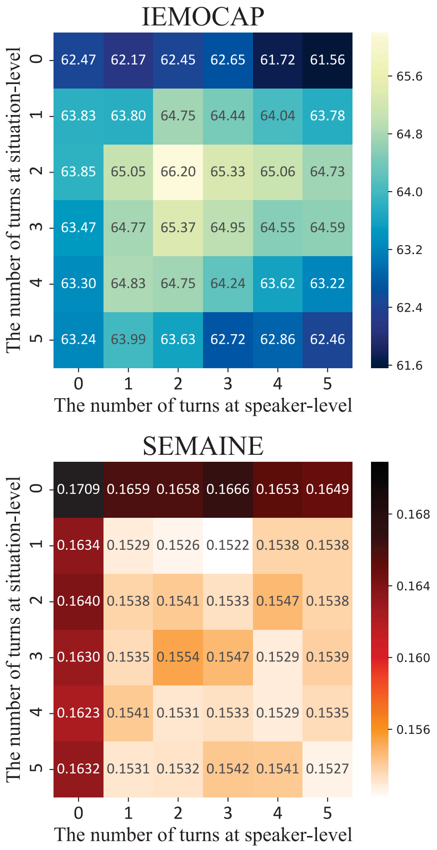
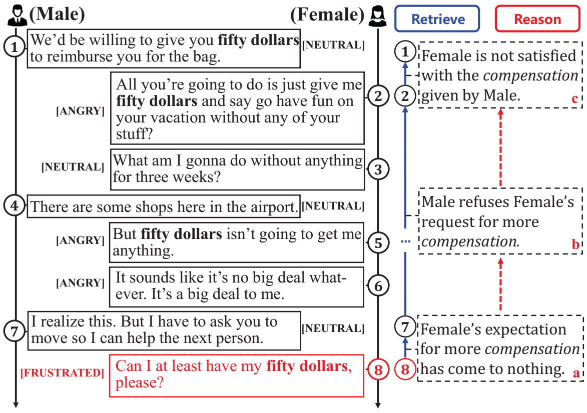


Figure 3: Results against the number of turns. We report the *Weighted-F*1 score on the *IEMOCAP* dataset and *MAE* of *Arousal* attribute on the *SEMAINE* dataset. The lighter the color, the better the performance.

4.3 Parameter Analysis

We investigate how our model performs w.r.t the number of turns in the cognitive phase. From Figure 3, the best $\{T^s, T^v\}$ is $\{2, 2\}$ and $\{1, 3\}$ on *IEMOCAP* and *SEMAINE* datasets, which obtain 66.20% *Weighted-F*1 and 0.1522 *MAE* of *Arousal* attribute, respectively. Note that the *SEMAINE* dataset needs more turns for the speaker-level cognitive phase. It implies speaker-level contextual clues may be more vital in arousal emotion, espe-



1 We'd be willing to give you **fifty dollars** to reimburse you for the bag. [NEUTRAL]

[ANGRY] All you're going to do is just give me **fifty dollars** and say go have fun on your vacation without any of your stuff? 2

[NEUTRAL] What am I gonna do without anything for three weeks? 3

4 There are some shops here in the airport. [NEUTRAL]

[ANGRY] But **fifty dollars** isn't going to get me anything. 5

[ANGRY] It sounds like it's no big deal whatever. It's a big deal to me. 6

7 I realize this. But I have to ask you to move so I can help the next person. [NEUTRAL]

[FRUSTRATED] Can I at least have my **fifty dollars**, please? 8

Female is not satisfied with the *compensation* given by Male. c

Male refuses Female's request for more *compensation*. b

Female's expectation for more *compensation* has come to nothing. a

Figure 4: The case study.

cially empathetic clues that require complex reasoning.

Besides, if we solely consider either situation-level or speaker-level context in the cognitive phase, results on the two datasets are significantly improved within a certain number of turns. The fact indicates the effectiveness of using multi-turn reasoning modules to understand contextual clues.

4.4 Case Study

Figure 4 shows a conversation sampled from the *IEMOCAP* dataset. The goal is to predict the emotion label of *utterance* 8. Methods such as DialogueRNN and DialogueGCN lack the ability to consciously understand emotional clues, *e.g.*, the cause of the emotion (failed expectation). They are easy to mistakenly identify the emotion as *angry* or *neutral*.

Our model DialogueCRN can understand the conversational context from a cognitive perspective. In the cognitive phase, the following two processes are performed iteratively: the intuitive retrieving process of *8-7-2-1* (blue arrows) and the conscious reasoning process of *a-b-c* (red arrows), to extract and integrate emotional clues. We can obtain that *utterance* 8 implied that more compensation expected by *female* was not achieved. The failed compensation leads to more negative of his emotion and thus correctly identified as *depression*.

5 Related Work

5.1 Emotion Recognition

Emotion recognition (ER) has been drawing increasing attention to natural language processing (NLP) and artificial intelligence (AI). Existing works generally regard the ER task as a classification task based on context-free blocks of data, such as individual reviews or documents. They can roughly divided into two parts, *i.e.*, feature-engineering based (Devillers and Vidrascu, 2006), and deep-learning based methods (Tang et al., 2016; Wei et al., 2020).

5.2 Emotion Recognition in Conversations

Recently, the task of Emotion Recognition in Conversations (ERC) has received attention from researchers. Different traditional emotion recognition, both situation-level and speaker-level context plays a significant role in identifying the emotion of an utterance in conversations (Li et al., 2020). The neglect of them would lead to quite limited performance (Bertero et al., 2016). Existing works generally capture contextual characteristics for the ERC task by deep learning methods, which can be divided into *sequence-based* and *graph-based* methods.

Sequence-based Methods. Many works capture contextual information in utterance sequences. Poria et al. (2017) employed LSTM (Hochreiter and Schmidhuber, 1997) to capture conversational context features. Hazarika et al. (2018a,b) used end-to-end memory networks (Sukhbaatar et al., 2015) to capture contextual features that distinguish different speakers. Zhong et al. (2019); Li et al. (2020) utilized the transformer (Vaswani et al., 2017) to capture richer contextual features based on the attention mechanism. Majumder et al. (2019) introduced a speaker state and global state for each conversation based on GRUs (Cho et al., 2014). Moreover, Jiao et al. (2020a) introduced a conversation completion task to learn from unsupervised conversation data. Jiao et al. (2020b) proposed a hierarchical memory network for real-time emotion recognition without future context. Wang et al. (2020) modeled ERC as sequence tagging to learn the emotional consistency. Lu et al. (2020) proposed an iterative emotion interaction network to explicitly model the emotion interaction.

Graph-based Methods. Some works (Zhang et al., 2019; Ghosal et al., 2019; Ishiwatari et al., 2020; Lian et al., 2020) model the conversational context by designing a specific graphical structure. They utilize graph neural networks (Kipf and Welling, 2017; Velickovic et al., 2017) to capture multiple dependencies in the conversation, which have achieved appreciable performance.

Different from previous works, inspired by the *Cognitive Theory of Emotion* (Schachter and

Singer, 1962; Scherer et al., 2001), this paper makes the first attempt to explore cognitive factors for emotion recognition in conversations. To sufficiently understand the conversational context, we propose a novel DialogueCRN to extract and then integrate rich emotional clues in a cognitive manner.

6 Conclusion

This paper has investigated cognitive factors for the task of emotion recognition in conversations (ERC). We propose novel contextual reasoning networks (DialogueCRN) to sufficiently understand both situation-level and speaker-level context. DialogueCRN introduces the cognitive phase to extract and integrate emotional clues from context retrieved by the perceptive phase. In the cognitive phase, we design multi-turn reasoning modules to iteratively perform the intuitive retrieving process and conscious reasoning process, which imitates human unique cognitive thinking. Finally, emotional clues that trigger the current emotion are successfully obtained and used for better classification. Experiments on three benchmark datasets have proved the effectiveness and superiority of the proposed model. The case study shows that considering cognitive factors can better understand emotional clues and boost the performance of ERC.

References

Alan Baddeley. 1992. Working memory. *Science*, 255(5044):556–559.

Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. 2016. Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *EMNLP*, pages 1042–1047. The Association for Computational Linguistics.

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.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734. The Association for Computer Linguistics.

Laurence Devillers and Laurence Vidrascu. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *INTERSPEECH*. ISCA.

Jonathan St BT Evans. 1984. Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75(4):451–468.

Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.

Jonathan St BT Evans. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59:255–278.

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 *EMNLP/IJCNLP (1)*, pages 154–164. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: interactive conversational memory network for multimodal emotion detection. In *EMNLP*, 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 *NAACL-HLT*, pages 2122–2132. Association for Computational Linguistics.

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

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 *LREC*. European Language Resources Association (ELRA).

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *EMNLP (1)*, pages 7360–7370. Association for Computational Linguistics.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020a. Exploiting unsupervised data for emotion recognition in conversations. In *EMNLP (Findings)*, pages 4839–4846. Association for Computational Linguistics.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020b. Real-time emotion recognition via attention gated hierarchical memory network. In *AAAI*, pages 8002–8009. AAAI Press.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751. The Association for Computer Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR (Poster)*. OpenReview.net.

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 *PervasiveHealth*, pages 206–210. ACM.

Akshi Kumar, Prakhar Dogra, and Vikrant Dabas. 2015. Emotion analysis of twitter using opinion mining. In *IC3*, pages 285–290. IEEE Computer Society.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *COLING*, pages 4190–4200. International Committee on Computational Linguistics.

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.

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 *COLING*, 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 *AAAI*, pages 6818–6825. AAAI Press.

Gary McKeown, Michel François Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.*, 3(1):5–17.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814. Omnipress.

Jérémie Nicolle, Vincent Rapp, Kevin Bailly, Lionel Prevost, and Mohamed Chetouani. 2012. Robust continuous prediction of human emotions using multiscale dynamic cues. In *ICMI*, pages 501–508. ACM.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. The Association for Computer Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL (1)*, pages 873–883. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL (1)*, pages 527–536. Association for Computational Linguistics.

Francisco A. Pujol, Higinio Mora, and Ana Martínez. 2019. Emotion recognition to improve e-healthcare systems in smart cities. In *RIIFORUM*, pages 245–254. Springer.

Stanley Schachter and Jerome Singer. 1962. Cognitive, social and physiological determinants of emotional state. *Psychological Review*, 69:378–399.

Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Björn W. Schuller, Michel François Valstar, Roddy Cowie, and Maja Pantic. 2012. AVEC 2012: the continuous audio/visual emotion challenge - an introduction. In *ICMI*, pages 361–362. ACM.

Steven A Sloman. 1996. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *NIPS*, pages 2440–2448.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In *COLING*, pages 3298–3307. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks. In *ICLR*.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *SIGdial*, pages 186–195. Association for Computational Linguistics.

Lingwei Wei, Dou Hu, Wei Zhou, Xuehai Tang, Xiaodan Zhang, Xin Wang, Jizhong Han, and Songlin Hu. 2020. Hierarchical interaction networks with rethinking mechanism for document-level sentiment analysis. In *ECML/PKDD*.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421. ijcai.org.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *EMNLP/IJCNLP (1)*, pages 165–176. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguistics*, 46(1):53–93.