CAN-GRU: a Hierarchical Model for Emotion Recognition in Dialogue

Ting Jiang **Bing Xu** **Tiejun Zhao** **Sheng Li**
Laboratory of Machine Intelligence and Translation
School of Computer Science and Technology, Harbin Institute of Technology
Harbin, China
jiangting_hit@163.com {hitxb,tjzhao,lisheng}@hit.edu.cn

Abstract

Emotion recognition in dialogue systems has gained attention in the field of natural language processing recent years, because it can be applied in opinion mining from public conversational data on social media. In this paper, we propose a hierarchical model to recognize emotions in the dialogue. In the first layer, in order to extract textual features of utterances, we propose a convolutional self-attention network(CAN). Convolution is used to capture n-gram information and attention mechanism is used to obtain the relevant semantic information among words in the utterance. In the second layer, a GRU-based network helps to capture contextual information in the conversation. Furthermore, we discuss the effects of unidirectional and bidirectional networks. We conduct experiments on Friends dataset and EmotionPush dataset. The results show that our proposed model(CAN-GRU) and its variants achieve better performance than baselines.

1 Introduction

As an important component of human intelligence, emotional intelligence is defined as the ability to perceive, integrate, understand and regulate emotions(Mayer et al., 2008). Emotion is the essential difference between human and machine, so emotion understanding is an important research direction of artificial intelligence. As the most common way for people to communicate in daily life, dialogue contains a wealth of emotions. Recognising the emotions in the conversation is of great significance in intelligent customer service, medical systems, education systems and other aspects.

According to (Poria et al., 2015), textual features usually contain more emotional information than video or audio features, so we focus on the emotion analysis of dialogue text and aims to recognize the emotion of each utterance in dialogues.

There are some challenges in this task. First, the length of an utterance may be too long, making it difficult to capture contextual information. Furthermore, a dialogue usually contains lots of utterances, therefore, it's hard to grasp long-term contextual relations between utterances. Second, the same word may express different emotions in different contexts. For example, in Table 1, while in different dialogues, the word 'Yeah' can express three different emotions, that is , joy, neutral and suprise. To tackle these challenges, we propose a hierarchical model based on convolutional attention network and gated recurrent unit (CAN-GRU). Existing works pay little attention to the extraction of semantic information within an utterance. In this work, we focus on this problem. Our proposed model can extract n-gram information by CNNs and use self-attention to capture contextual information within an utterance in the first layer. Moreover, we utilize a GRU-based network to model the sequence of utterances in the second layer, which can fully combine the context when analyzing utterance emotion and solve the problem of long-term dependence between texts at the same time.

2 Related Work

Text emotion recognition is one of the most hot topic in natural language processing. Recent years,a lot of classical neural networks are used to tackle this problem. Such as Long Short-Term Memory Net-

speaker	utterance	emotion
Phoebe	Can I tell you a little secret?	neutral
Rachel	**Yeah!**	**joy**
Wayne	Hey Joey, I want to talk to you.	neutral
Joey	**Yeah?**	**neutral**
Gary	Hey Chandler, what are you doing here?	suprise
Chandler	Gary, I'm here to report a crime.	neutral
Gary	**Yeah?**	**suprise**
Chandler	It is a crime that you and I don't spend more time together.	neutral

Table 1: The word 'Yeah' expresses different emotions in the different contexts.

work(Hochreiter and Schmidhuber, 1997), Gated Recurrent Unit Network(Cho et al., 2014) and textual Convolutional Neural Network(Kim, 2014). However, these models don't perform well when the texts are too long, because it's hard to capture the long-range contextual information. Later, attention mechanism(Bahdanau et al., 2015) is proposed to solve this problem. Recently, self-attention(Vaswani et al., 2017) is widely used since it can solve the long-term dependence problem of text effectively.

Recent years, more and more researchers focus on emotion recognition in conversation. This task aims to recognize the emotion of each utterance in dialogues. bcLSTM(Poria et al., 2017) extracts textual features by CNN and model the sequence of utterances by LSTM. Considering inter-speaker dependency relations, conversational memory network(CMN)(Hazarika et al., 2018b) has been proposed to model the speaker-based emotion using memory network and summarize task-specific details by attention mechanisms. ICON(Hazarika et al., 2018a) improves the CMN, it hierarchically models the self-speaker emotion and inter-speaker emotion into global memories. DialogueRNN(Majumder et al., 2019) uses emotion GRU and global GRU to model inter-party relation, and uses party GRU to model relation between two sequential states of the same party. DialogueGCN(Ghosal et al., 2019) improves DialogueRNN by graph convolutional network, and it can hold richer context relevant to emotion. However, these models may be too complex for small textual dialogue datasets.

In this paper, we study on the EmotionX Challenge(Hsu and Ku, 2018), Dialogue Emotion Recognition Challenge, which aims to recognize the emotion of each utterance in dialogues. According to the overview of this task, the best team(Khosla, 2018) proposes a CNN-DCNN auto encoder based model, which includes a convolutional encoder and a deconvolutional decoder. The second place team(Luo et al., 2018) mainly uses BiLSTM with a self-attentive architecture on the top for the classiffication. The third place team(Saxena et al., 2018) proposes a hierarchical network based on attention models and conditional random fields(CRF). For a meaningful comparison, we use the same dataset and metric as the challenge in our study.

3 Method

3.1 Task Definition

Given a dialogue $dia = \{u_1, u_2, ..., u_N\}$, where N is the number of utterances in the dialogue, $u_i = \{w_1, w_2, ..., w_L\}$ represents the $ith(1 \leq i \leq N)$ utterance in the dialogue that consists of L words, our goal is to analyze the emotion of each utterance in the dialogue. To solve this task, we propose a hierarchical model CAN-GRU and extend three variants, CAN-GRUA, CAN-biGRU and CAN-biGRUA(illustrated in Fig. 1).

3.2 Text Feature Extraction

In this section, we discuss the first layer of the model. Like (Poria et al., 2017), we use convolutional neural network to extract the features of the utterance. Inspired by (Gao et al., 2018), in order to capture the contextual information of long text effectively, we use convolutional self-attention network(CAN) instead of traditional CNN network.

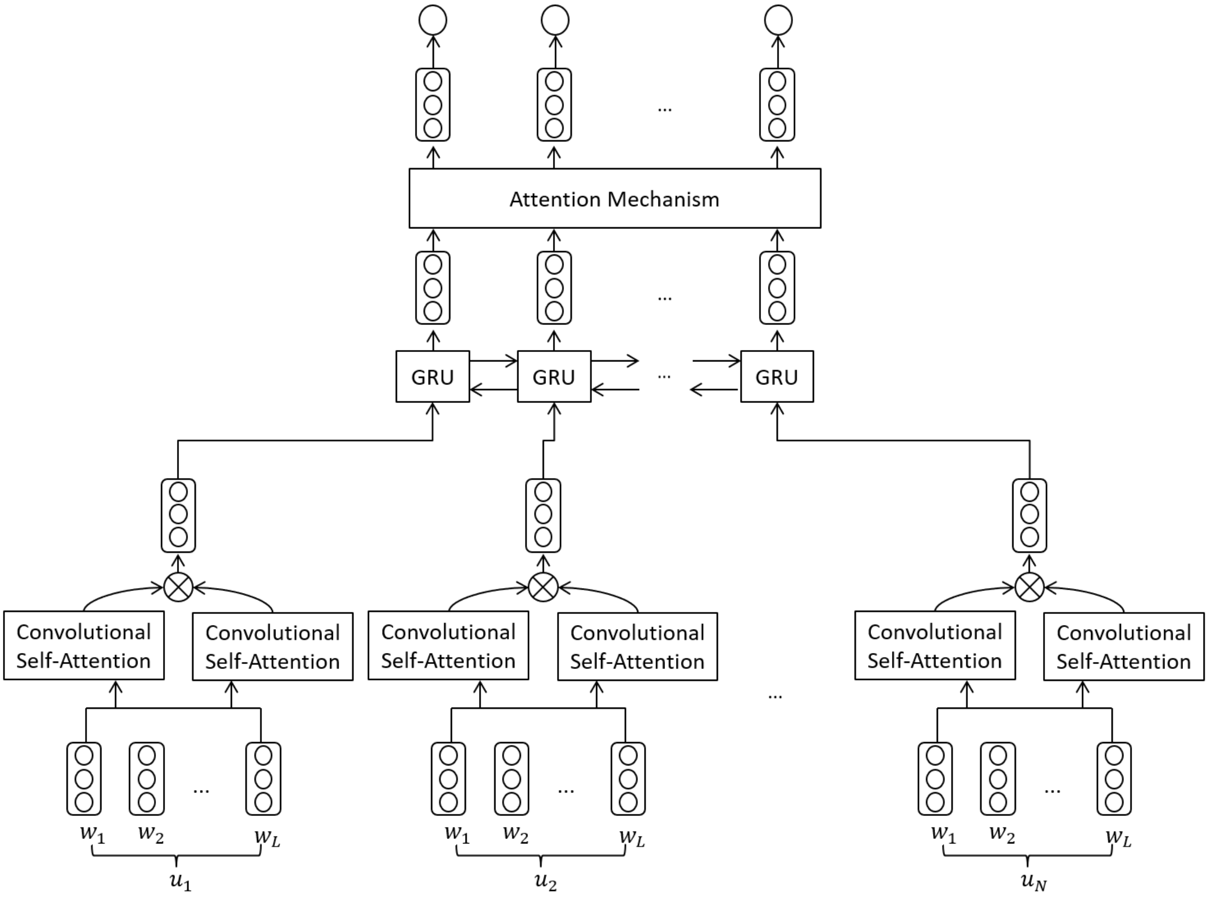

Attention Mechanism
GRU
GRU
GRU
Convolutional Self-Attention
Convolutional Self-Attention
Convolutional Self-Attention
Convolutional Self-Attention
Convolutional Self-Attention
Convolutional Self-Attention
w_1 w_2 w_L
u_1
w_1 w_2 w_L
u_2
w_1 w_2 w_L
u_N

Figure 1: The architecture of our proposed CAN-biGRUA. In the first layer, convolutional neural network and self-attention mechanism are used to extract text features. In the second layer, biGRU with an attentive architecture on the top is used to model the sequence of utterances in the dialogue.

For query embedding Q, key embedding K and value embedding V involved in attention operation, they may need different effective features. And different effective information can be extracted in different convolution operations. So we obtain the Q, K and V embeddings by convolving the input word embeddings, instead of using the input word embeddings as the Q, K and V embeddings directly:

$$Q = f(conv(E, W_q) + b_q) \tag{1}$$

$$K = f(conv(E, W_k) + b_k) \tag{2}$$

$$V = f(conv(E, W_v) + b_v) \tag{3}$$

In the equations above, E is the input word embeddings, $\{E, Q, K, V\} \in \mathbb{R}^{l \times d}$, where l means the length of the sentence and d means the embedding dimension. $\{W_q, W_k, W_v\} \in \mathbb{R}^{w \times n \times d}$, where w is the window size of filters and n is the feature maps of filters. $\{b_q, b_k, b_v\} \in \mathbb{R}^d$. $conv(E, W)$ means convolution operation between E and W. And f is the activation function.

After getting Q, K, V embeddings, we calculate semantic relations among words within the utterance by the scaled dot product attention operation. More specificly, Q and K operate to get the weight matrix. Then we scale this weight matix by $\sqrt{d}$. After that, softmax operation is conducted to obtain the standardized weight matrix, which is used to express the degree of attention between words in the sentence, then the normalized weight matrix is mutiplied with V to get the result $Z \in \mathbb{R}^{l \times d}$ of attention operation:

$$Z = softmax(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

As mentioned in (Gao et al., 2018), attention mechanisms cannot capture complex interactions, because it is designed for creating weighted averages. So we do the equations(1)-(3) twice to create Q_a, K_a, V_a and Q_b, K_b, V_b respectively, and get Z_a, Z_b by operate equation(4) respectively, then perform elementwise multiplication on Z_a and Z_b to get $U \in \mathbb{R}^{l \times d}$:

$$U = Z_a \otimes Z_b \tag{5}$$

Finally, for each $U_i(1 \leq i \leq N)$ in $dia = \{U_1, U_2, ..., U_N\}$, we get the individual embedding e_{u_i} by max-pooling on the contextual word embeddings within the U_i. In this way, we obtain a set of utterance embeddings $\{e_{u_1}, e_{u_2}, ..., e_{u_N}\}$ in one dialogue.

3.3 Dialogue Modeling

In this section, we discuss the sencond layer of the model. Considering different networks, we propose the hierarchical model CAN-GRU and its three progressive variants.

CAN-GRU: In real life, when we analyze the emotion of the current utterance, we can only refer to the historical information of the past utterances in the conversation. So in our model, we use GRU to model the sequence of utterances in the dialogue, because it can memory and transmit historical information. GRU(Cho et al., 2014) is an improved model for the original recurrent neural networks and it performs well with simple calculation. At timestep t, it use reset gate R_t and update gate Z_t to calculate current hidden state S_t with input utterance embedding e_{u_t} and hidden state s_{t-1} at the previous time step.

$$R_t = \sigma(e_{u_t} W_{ur} + S_{t-1} W_{sr} + b_r) \tag{6}$$

$$Z_t = \sigma(e_{u_t} W_{uz} + S_{t-1} W_{sz} + b_z) \tag{7}$$

$$H_t = tanh(e_{u_t} W_{uh} + (R_t \otimes S_{t-1}) W_{sh} + b_h) \tag{8}$$

$$S_t = Z_t \otimes S_{t-1} + (1 - Z_t) \otimes H_t \tag{9}$$

where W, b are trainable parameters and $\otimes$ means elementwise mutilication.

CAN-GRUA: However, it is difficult to grasp long-term dependence between sentences when there are too many sentences in a conversation. That is, it is hard for the current utterance to capture the historical information contained in the distant utterance. To solve this problem, we connect an attention layer upon the GRU to obtain the influence degree of historical information on the emotion of the current utterance. If the weight calculated by attention mechanism tends to be large, it indicates that the preceding utterance have an important influence on the current utterance, so this preceding utterance should be given more attention.

$$\tilde{S_t} = \sum_{i=1}^{t-1} S_i \alpha_i \tag{10}$$

$$\alpha_i = \frac{exp(S_t S_i)}{\sum\limits_{i=1}^{t-1} exp(S_t S_i)} \tag{11}$$

Here, $S_t \in \mathbb{R}^m$ is the current hidden state, where m is the dimension of the hidden state. $S_i \in \mathbb{R}^m$ is the preceding hidden state at time step i, $\tilde{S_t} \in \mathbb{R}^m$ is the attention result at time step t.

CAN-biGRU: In fact, when analyzing the emotion of the utterance, we can not only use the historical information before the utterance, but also the future information after the current utterance. This is because emotional tone is usually maintained and does not shift frequently within a conversation in a short time. If we only pay attention to the historical information, it may be difficult to analyze the emotion of the current utterance, while the future information can be helpful in the analysis. Therefore, using both historical and future information can help to capture a richer context. Bidirectional GRU(biGRU) is used to model the sequence of utterances abstracting contextual features forward and backward, which can provides context for emotion classification more effectively.

CAN-biGRUA: As mentioned before, biGRU also suffers from the difficulty of obtaining semantic connections between long sequences. So we connect a self-attention layer on the top of the hidden states of biGRU to take full advantage of global contextual information.

$$\tilde{S_t} = \sum_{i=1}^{N} S_i \alpha_i \tag{12}$$

$$\alpha_i = \frac{exp(S_t S_i)}{\sum\limits_{i=1}^{N} exp(S_t S_i)} \tag{13}$$

Here, $S_t \in \mathbb{R}^m$ is the current hidden state, $S_i \in \mathbb{R}^m$ is the hidden state at time step i, N is the number of sentences in the dialogue, $\tilde{S_t} \in \mathbb{R}^m$ is the attention result at time step t.

3.4 Emotion Classification

As mentioned above, we get final representations of utterances $\{\tilde{S_1}, \tilde{S_2}, ..., \tilde{S_t}, ..., \tilde{S_N}\}$. Then we utilize a fully-connected layer and a softmax layer to get the emotion class of each utterance in a dialogue.

$$f_t = tanh(W_f \tilde{S_t} + b_f) \tag{14}$$

$$o_t = softmax(W_o \tilde{f_t} + b_o) \tag{15}$$

$$\hat{y_t} = \underset{i}{argmax}(o_t[i]), i \in [1, c] \tag{16}$$

Where $W_f \in \mathbb{R}^{m \times m}$, $b_f \in \mathbb{R}^m$. $W_o \in \mathbb{R}^{m \times c}$, c is the number of emotion class, $b_o \in \mathbb{R}^c$, $o_t \in \mathbb{R}^c$, $\hat{y_t}$ is the predicted class for utterance u_t.

3.5 Training

Like(Khosla, 2018), in order to solve the problem of emotion class imbalance, we use a weigted cross entropy loss as a minimization target to optimize the parameters in the model. We give higher weight to the loss of minority class data sample in the dataset.

$$Loss = \frac{1}{K} \sum_{k=1}^{K} weight_k loss_k \tag{17}$$

$$loss_k = -[y_k log(p_k) + (1 - y_k) log(1 - p_k)] \tag{18}$$

$$\frac{1}{weight_k} = \frac{count_i}{\sum\limits_{i=1}^{c} count_i} \tag{19}$$

Where K is the total number of samples, y_k is the ground-truth, p_k is the probability calculated in softmax layer, $count_i$ is the total number of samples in the same class as sample k.

4 Experiments

4.1 Datasets

We conduct experiments on two datasets provided by the EmotionX Challenge(Hsu and Ku, 2018).

Friends[0] : The conversations in this dataset are from the Friends TV show transcripts. The dataset contains eight emotion categories: joy, anger, sadness, surprise, fear, disgust, neutral, and non-neutral.

EmotionPush[1] : The conversations in this dataset are from the facebook messenger logs after processing the private information. Emotion categories are the same as Friends dataset.

[0] http://doraemon.iis.sinica.edu.tw/emotionlines

[1] http://doraemon.iis.sinica.edu.tw/emotionlines

In the challenge(Hsu and Ku, 2018), each dataset is divided into the training set with 720 dilogues, the validation set with 80 dialogues and the test set with 200 dialogues. Since there are few utterances for some emotions, the challenge only evaluate the performence of recognition for four emotions: joy, anger, sadness and neutral. Table 2 shows the distributions of train, validation, test samples and the distributions of the emotions for both datasets respectively.

dataset	Dialogue(Utterance)			Emotion				
	train	validation	test	anger	joy	sadness	neutral	others
Friend	720(10561)	80(1178)	200(2764)	759	1710	498	6530	5006
EmotionPush	720 (10,733)	80(1202)	200(2807)	140	2100	514	9855	2133

Table 2: Statistics of the datasets.

4.2 Evaluation Metric

We use the unweighted accuracy(UWA) as the evaluation metric instead of the weighted accuracy(WA), the same as the challenge. This is because WA is easily influced by the large proportion of neutral emotion and UWA can help to make a meaningful comparision.

$$UWA = \frac{1}{c}\sum_{i=1}^{c} a_i, WA = \sum_{i=1}^{c} weight_i a_i \tag{20}$$

Where a_i is the accuracy of class i and $weight_i$ is the percentage of the class i.

4.3 Experimental Setting

We use 300-dimensional pre-trained GloVe[2] (Pennington et al., 2014) word-embeddings which is trained from web data. We use three distinct convolution filters of sizes 3, 4, and 5 respectively, each having 100 feature maps. The dimension of the hidden states of the GRU is set to 300. We use adam(Kingma and Ba, 2015) optimizer and set the initial learning rate as 1.0×10^{-4}. The learning rate is halved every 20 epochs during training. Dropout probability is set to 0.3.

4.4 Baselines

In experiments, we compare our proposed model with the following models.

CNN-DCNN: The winner of EmotionX Challenge(Khosla, 2018). The model contains a convolutional encoder and a deconvolutional decoder. The linguistic features enhance the latent feature of the model.

SA-LSTM: The second place of the challenge(Luo et al., 2018). A self-attentive biLSTM network can provide information between utterances and the word dependency in each utterance.

HAN: The third place of the challenge(Saxena et al., 2018). LSTM with attention mechanism gets the sentence embedding. Another LSTM and CRF layer model the context dependency between sentence embeddings of the dialogue.

scGRU: We implement the basic model proposed by(Poria et al., 2017), but with a few changes. The same as (Poria et al., 2017), CNN is used to extract text features, but we use a contextual GRU network instead of a contextual LSTM network to model the sequences.

bcGRU: We implement the variant model proposed by(Poria et al., 2017), CNN is also used to obtain utterance features, but the biLSTM network used in the author's work is replaced by biGRU network.

4.5 Main Results

Table 3 presents the performance of baselines and CAN-GRU along with its variants.

Baselines: Our implemented bcGRU model performes better than scGRU on both datasets. On the Emotionpush dataset, bcGRU's performance has surpassed CNN-DCNN, and it is the best model in baselines. On the Friend dataset, CNN-DCNN remains the best baseline.

[2] http://nlp.stanford.edu/projects/glove/

model	Friend					EmotionPush				
	anger	joy	sadness	neutral	UWA	anger	joy	sadness	neutral	UWA
CNN-DCNN	55.3	71.1	55.3	68.3	62.5	45.9	**76.0**	51.7	76.3	62.5
SA-BiLSTM	49.1	68.8	30.6	**90.1**	59.6	24.3	70.5	31.0	**94.2**	55.0
HAN	39.8	57.6	50.6	73.5	55.4	21.6	63.1	54.0	88.2	56.7
scGRU	51.6	68.7	44.8	72.6	59.4	49.8	68.2	57.1	75.4	62.6
bcGRU	54.1	69.8	43.5	73.4	60.2	50.1	71.4	61.6	71.8	63.7
CAN-GRU	56.2	67.0	**55.9**	71.4	62.6	52.4	70.6	59.8	74.5	64.3
CAN-biGRU	54.8	68.1	52.9	76.3	63.0	**55.7**	71.8	60.1	74.9	65.6
CAN-GRUA	**57.6**	70.2	53.7	76.2	64.4	53.2	72.1	61.5	78.3	66.3
CAN-biGRUA	56.4	**72.6**	54.4	77.8	**65.3**	54.3	73.8	**62.9**	77.4	**67.1**

Table 3: Experimental results on Friend dataset and EmotionPush dataset.

CAN-GRU: In the first layer, it uses the convolutional self-attention mechanism to extract utterance features, and in the second layer, GRU is used to model the sequence of utterances. Compared with scGRU, it attains 3.2% and 1.7% improvement on the Friend dataset and EmotionPush dataset.

CAN-biGRU: Compared with CAN-GRU, it uses biGRU at the second layer and get improvements on the two datasets. CAN-biGRU achieves 2.8% and 1.9% improvements over bcGRU on the Friend dataset and the Emotionpush dataset respectively. Both the improvements of CAN-GRU and CAN-biGRU over baselines illustrate that the convolutional self-attention mechanism can capture contextual information in long text effectively.

CAN-GRUA: Compared with CAN-GRU, an attention mechanism is connected upon the GRU layer, which can help the model better capture the historical information of utterance and give high weight to important historical information. It gets 1.8% and 2.0% improvements over CAN-GRU on the two datasets.

CAN-biGRUA: At the top of biGRU, a self-attention mechanism is added to help calculate the importance of contextual information by using historical and future information when analyzing the current utterance emotion. This model achieves the best results, it improves 2.8% and 4.6% over baseline on the two datasets respectively.

model	Friend					EmotionPush				
	anger	joy	sadness	neutral	UWA	anger	joy	sadness	neutral	UWA
BERT	78.1	86.5	74.3	**90.3**	82.3	79.4	**89.7**	85.3	92.4	86.7
CAN-biGRU(*)	**81.2**	**87.4**	**78.7**	89.1	**84.1**	**82.8**	88.3	**87.6**	**94.1**	**88.2**

Table 4: Experimental results for BERT and CAN-biGRU(*).

In addition, we use the pretrained model BERT(Devlin et al., 2019) to get the word embeddings and input the pre-trained word embeddings into our CAN-biGRU, the experimental results are shown as the CAN-biGRU(*) in Table 4. As we can see, while BERT achieves a high degree of accuracy, our model can be further improved on the basis of BERT. CAN-biGRU(*) gets 1.8% and 1.5% improvements over BERT on the Friend dataset and the Emotionpush dataset respectively.

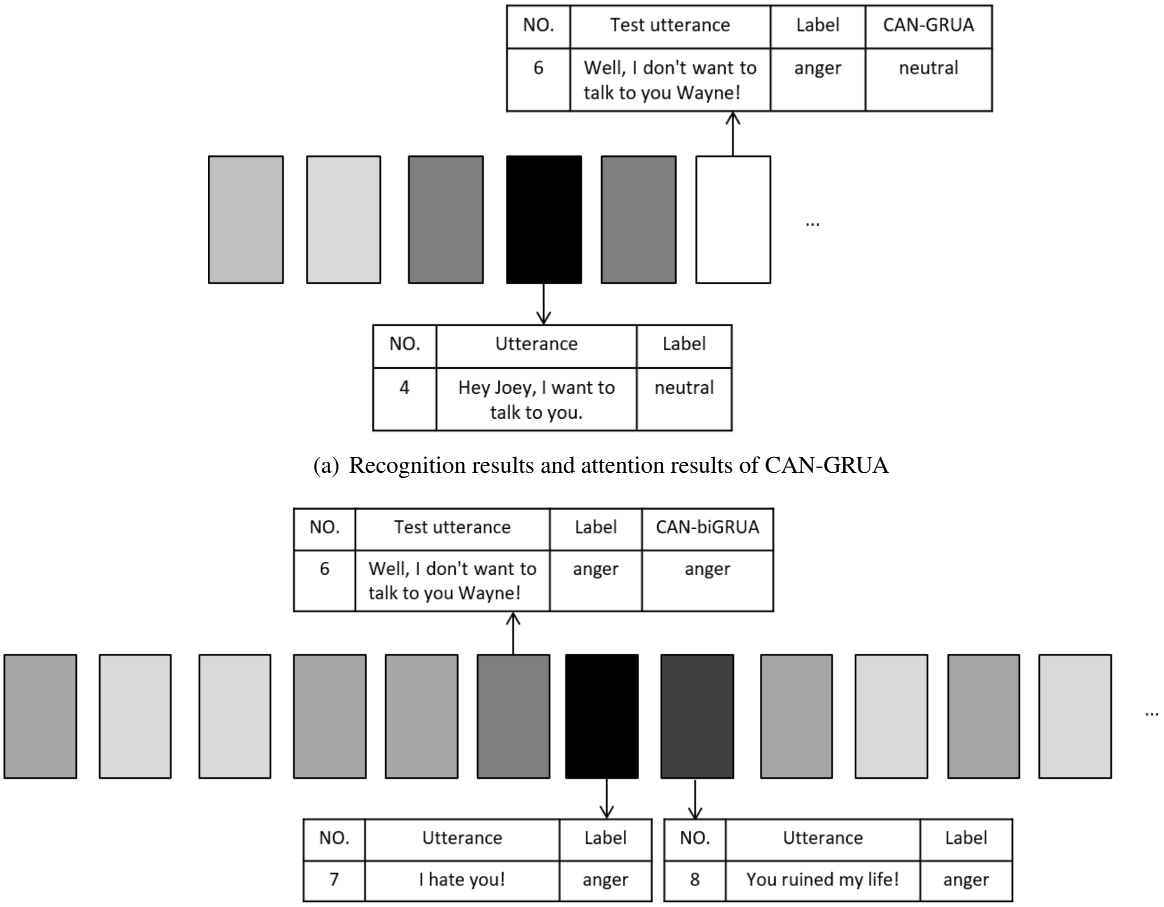
4.6 Case Study

In Table 5, we compare the emotion recognition results of bcGRU and CAN-biGRU. In the first case, ‘bad’ expresses strong emotion and both two model can recognize the sad emotion successfully. While there is no explicit emotion word in the third utterance, but the word ‘ruined’ delivers bad information, our CAN-biGRU can extract semantic information among words by CAN and gives the right prediction. In the second case, the word ‘celebrating’ in the third utterance express the joy emotion implicitly. Our model obtains the contextual information through the CAN, and makes the correct prediction. However,

speaker	utterance	True label	bcGRU	CAN-biGRU
Phoebe	Oh, it's bad. It's really bad ... Which I do.	sadness	**sadness**	**sadness**
Chandler	How's your room Rach?	neutral	**neutral**	**neutral**
Rachel	Everything's ruined ... blue sweater.	sadness	neutral	**sadness**
Joey	Hey-hey-hey!	joy	**joy**	**joy**
Chandler	What are you doing?	neutral	**neutral**	**neutral**
Phoebe	We're just celebrating that Joey ... back.	joy	neutral	**joy**
Phoebe	I'm sorry ... Check this out.	neutral	sad	sad
Monica	No, Phoebe ... you play it at the wedding.	neutral	**neutral**	**neutral**

Table 5: Some case comparisons of emotion recognition results by bcGRU and CAN-biGRUA

in the third case, both two model make false predictions for the utterance said by Phoebe, since the word 'sorry' expresses strong sad emotion. This shows CAN is still limited in such complicated semantic environment.



NO.	Test utterance	Label	CAN-GRUA
6	Well, I don't want to talk to you Wayne!	anger	neutral

NO.	Utterance	Label
4	Hey Joey, I want to talk to you.	neutral

(a) Recognition results and attention results of CAN-GRUA

NO.	Test utterance	Label	CAN-biGRUA
6	Well, I don't want to talk to you Wayne!	anger	anger

NO.	Utterance	Label
7	I hate you!	anger

NO.	Utterance	Label
8	You ruined my life!	anger

(b) Recognition results and attention results of CAN-biGRUA

Figure 2: Comparison of recognition results and attention results between CAN-GRUA and CAN-biGRUA. Deeper color means higher attention.

As shown in Table 6, we analyse some cases of the results of emotion recognition by our CAN-biGRUA. In the first two cases, our model can successfully recognise the emotion category of utterances. In the conversation of Monica and Joey, 'Yeah' expresses neutral emotion, while in the conversation of Chloe and Ross, 'Yeah' with '!' expresses stronger emotion and our model analyses its joy emotion effectively. However, in the third case, our model makes wrong classification for the first,third and fourth utterances.

This dialogue contains three different emotions and emotion shifts frequently. The failure of our model indicates that although considering the context, model's ability to understand emotions in the complicated situation is limited and still needs improvement.

In Fig. 2, we compare the recognition results and attention results of CAN-GRUA and CAN-biGRUA for the sixth utterance in the dialogue. As we can see, CAN-GRUA only uses historical information and focuses on the fourth utterance, and it takes neutral emotion as a result. While CAN-biGRUA takes both historical and future information into account, and it mainly pays attention to the seventh and the eighth utterances which contain strongly anger emotion, so the model finnaly classifies the test utterance as anger emotion. This case shows that considering both historical information and future information can help model make better classifications.

speaker	utterance	True label	Predicted label
Monica	Hey, Joey, could you pass the cheese?	**neutral**	**neutral**
Joey	Yeah.	**neutral**	**neutral**
Chloe	That's so great for you guys!	**joy**	**joy**
Ross	Yeah!	**joy**	**joy**
Chloe	Good luck, with your girlfriend.	**neutral**	**neutral**
Monica	Ross, we can handle this.	neutral	joy
Ross	Well,... be hurt over something that is so silly.	**sadness**	**sadness**
Ross	I mean, enough of the silliness!	anger	sadness
Chandler	Well, why don't you tell her to stop being silly!	anger	sadness

Table 6: Some cases of emotion recognition results by CAN-biGRUA

5 Conclusion

In the paper, we propose a hierarchical model(CAN-GRU) to tackle emotion recognition in dialogues. Unlike existing works, we focus on semantic information extraction within utterance in the dialogue. N-gram features and relevant semantic information among words in the utterance are learned by the convolutional self-attention network in the first layer and the sequence of utterances is modeled by the GRU-based network in the second layer. We improve CAN-GRU to three variants, CAN-biGRU, CAN-GRUA and CAN-biGRUA. Experimental results show that attention mechanism can help to grasp long-term dependency in the contexts effectively. CAN-biGRUA achieves better results than CAN-GRUA demonstrates that it is necessary to consider both past and future information of the utterance.In the future, we will try to explore deeper semantic information in the context and focus more on emotion shift to solve the problem of poor performance of the model in complex situations.

Acknowledgements

The work of this paper is funded by the project of National key research and development program of China (No. 2018YFC0830700).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.

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 *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers).

Shang Gao, Arvind Ramanathan, and Georgia D. Tourassi. 2018. Hierarchical convolutional attention networks for text classification. In *Proceedings of The Third Workshop on Representation Learning for NLP, Rep4NLP@ACL*, pages 11–23. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 154–164. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132. Association for Computational Linguistics.

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

Chao-Chun Hsu and Lun-Wei Ku. 2018. Socialnlp 2018 emotionx challenge overview: Recognizing emotions in dialogues. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL*, pages 27–31. Association for Computational Linguistics.

Sopan Khosla. 2018. Emotionx-ar: CNN-DCNN autoencoder based emotion classifier. In Lun-Wei Ku and Cheng-Te Li, editors, *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL*, pages 37–44. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751. ACL.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Linkai Luo, Haiqing Yang, and Francis Y. L. Chin. 2018. Emotionx-dlc: Self-attentive bilstm for detecting sequential emotions in dialogues. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL*, pages 32–36. Association for Computational Linguistics.

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

John D Mayer, Richard D Roberts, and Sigal G Barsade. 2008. Human abilities: emotional intelligence. *Annual review of psychology*, 59:507—536.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. ACL.

Soujanya Poria, Erik Cambria, and Alexander F. Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2539–2544. ACL.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 873–883. Association for Computational Linguistics.

Rohit Saxena, Savita Bhat, and Niranjan Pedanekar. 2018. Emotionx-area66: Predicting emotions in dialogues using hierarchical attention network with sequence labeling. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL*, pages 50–55. 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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5998–6008.