A Context-Aware Gated Recurrent Units with Self-Attention for Emotion Recognition

Lin Li$^{1,a}$, Miaomiao Yang$^{1,b,*}$

$^1$Chongqing University of Post and Telecommunications, Chongqing 400000, China
E-mail: $^a$lilin1@cqupt.edu.cn; $^b*$s181201007@stu.cqupt.edu.cn

Abstract. Text-oriented emotion recognition research has made significant research progress, however, very few works pay attention to learn the high-quality long-distance contextual information for the utterance emotion recognition. In this paper, we proposed a context-aware gated recurrent units with self-attention for emotion recognition. The two bidirectional gated recurrent units can obtain the sequence relationship between words and utterances. Compared with the self-attention mechanism that cannot capture long-distance contextual information, three contexts are used in the self-attention mechanism, namely global, deep and deep-global context, which can capture the high-quality long-distance contextual information. And the connection mechanism can enhance the embedding information of words and utterances. Experimental results on public Friends and EmotionPush datasets demonstrate that the three contexts in the self-attention mechanism outperforms several baselines on some emotion types recognition and indicate the effectiveness of the designed model, especially the deep-global context increased by 1% and 0.2% on UWA compared with all models.

1. Introduction

Emotion detection in the dialogues has attracted more and more researchers’ attention due to its wide range of applications, such as counseling [1], public opinion mining [2], financial forecasting [3], and intelligence like smart homes and chatbots [4]. Emotion recognition based on deep learning is mainly to classify the emotion of text.

According to the basic emotion theory proposed by Ekman and Friesen in 1971, emotion mainly includes joy, sadness, anger, fear, surprise, disgust and neutral [5], [6]. We communicate with others not only the semantic information of the utterance, but also the emotions conveyed by the utterance. In a dialogue emotion detection, its goal is to tag each utterance in a dialogue with the indicated emotion [7], which need to consider the long-distance contextual information. For example, as can be seen from Figure 1, such as a series of utterances collected from a Friends TV script between Rachel and Ross, the first utterance conveys sadness emotion from the words "terrible", "can’t", and "see". Similarly, the fifth utterance expresses the anger emotion through "don’t", "see" and "!"] symbol. The above needs to consider capturing the contextual information of words. The long-distance contextual information of utterances can be seen from the eighth utterance, which deliver joy emotion because of the sixth utterance.

In recent years, many ways have been proposed and achieved excellent performances for emotion recognition. Such as Golder et al. [8] used a sentiment dictionary-based method for sentiment analysis of Twitter message, but this method is ideal when the coverage of emotional words and the accuracy of labeling are high. Furthermore, like Rozgic et al. [9] proposed a binary SVM classifier, and it is a traditional machine learning method without contextual
information, but it cannot automatically extract features and obtain meaning between words. So, the above two emotion recognition methods are not suitable for emotion recognition of utterances in the dialogues because of the sparse semantics of utterances. With the widespread use of deep learning, to take advantage of context, a bidirectional contextual long short-term memory (LSTM) network, termed as bcLSTM [10], and the long short-term memory network with self-attention (SA-BiLSTM) model [11], which can capture the sequential relationship of utterances. However, these approaches cannot capture the long-distance contextual information of utterances. For capturing the long-distance contextual information of utterances, Jiao et al. [12] put forward a two-level gated recurrent unit structure with self-attention, but the self-attention mechanism are comparatively common and plainly. In addition, pre-trained context-dependent encoder for utterance-level emotion recognition is put forward by Jiao [13]. However, these methods could not capture the high-quality long-distance contextual information of words and utterances. Due to the context-aware self-attention mechanism [14] is performed well on the machine translation tasks, which makes our model more effective than other baseline models.

In this paper, we proposed a context-aware gated recurrent units with self-attention for emotion recognition, that can catch the high-quality long-distance contextual information in the dialogues. The bidirectional gated recurrent units gain knowledge the contextual word embedding and the contextual utterance embedding. Three contexts are used in the self-attention mechanism, namely global, deep and deep-global context, which can enhance the high-quality long-distance contextual information of words and utterances by capturing the overall meaning of the words and utterances, and the connection mechanism can enhance the embedding information of the words and utterances.

The rest of this article is organized as follows. In Section 2, we discuss the research methodology used, including contextual word embedding, contextual utterance embedding, self-attention mechanism and connection mechanism. Then, we describe experiment in Section 3, including datasets, evaluation, baselines and analysis of results. End in Section 4, make conclusion and plan future work.

2. Methodology

2.1. Contextual Word Embedding

In a dialogue task, the first task is to get the contextual information of words. Given a series of dialogues, it is composed of the three parts \( \{(u_i, s_i, a_i)\} \), where \( u_i \) is the utterance spoken by the speakers, and \( a_i \) is the emotion label of the utterance spoken by the speakers, and the Figure 2 shows the framework of our model.

![Figure 1. An illustration of emotion recognition of a series of utterances](image-url)
Figure 2. The framework of the propose model

We choose a bidirectional gated recurrent unit network, termed as BiGRU, to model the word sequence of an utterance. For the word embedding $e(w_j)$ of an utterance, we use the BiGRU to obtain the contextual word embedding $e_c(w_j)$, then an individual utterance embedding $e(u_i)$ is received through the max-pooling operation:

$$H_j = \begin{bmatrix} \overrightarrow{H_j} & \overleftarrow{H_j} \end{bmatrix}$$

$$e_c(w_j) = \tanh (W_v \cdot H_j + l_u)$$

$$e(u_i) = \text{maxpool}\left\{\{e_c(w_j)\}\right\}$$

where $\overrightarrow{H_j}$ is the forward GRU and $\overleftarrow{H_j}$ is the backward GRU, which is of great significance of contextual information of an utterance. Respectively, $w_v$ is the shape of $d_1 \times 2d_1$ and $l_u$ is the size of $d_1$, and $d_1$ is the scale of the BiGRU.

2.2. Contextual Utterance Embedding
If given a different context for an utterance, it can deliver different meanings just like a word and its context. So we employ additional BiGRU to capture the utterances sequence of the dialogues, which can gain the contextual utterance embedding:

$$H_i = \begin{bmatrix} \overrightarrow{H_i} & \overleftarrow{H_i} \end{bmatrix}$$

$$e_c(u_i) = \tanh (W_q \cdot H_i + l_w)$$

where $\overrightarrow{H_i}$ is the forward GRU and $\overleftarrow{H_i}$ is the backward GRU. Besides, $W_q$ is the shape of $d_2 \times 2d_2$ and $l_w$ is the size of $d_2$, and $d_2$ is the scale of the dialogue BiGRU. Then, the contextual utterance
\[ e_c(u_i) \text{ is sent directly to the fully connected layer and a softmax function is used to predict the emotion of each group of dialogues:} \]

\[ \hat{y}_j = \text{softmax} \left( W_o \cdot e_c(u_i) + b_o \right) \]  

where \( \hat{y}_j \) is the predicted vector of emotion category.

### 2.3. Self-attention Mechanism

Compared with the self-attention mechanism that cannot capture the highquality long-distance contextual information, we used three contexts in the self-attention mechanism, namely global, deep and deep-global context, to enhance the contextual information of words and utterances by capturing the overall meaning of the utterances. For the hidden layers of the two BiGRUs \( H = \{H_1, H_2, \ldots, H_n\} \), we can obtain three context vectors by operating on the hidden layers.

The first is to average the hidden layers to obtain the global context \( C \), which can enhance the contextual relationship between words and utterances by calculating the weight and capturing the meaning and syntax information of the words and utterances from horizontal:

\[ C = \bar{H} \]  

the deep context \( C \) is described as hidden layers of the BiGRU, which can capture the highquality long-distance information of words and utterances by integrating different types of semantic information of different layers from the vertical integration:

\[ C = H \]  

the deep-global context \( C \) is a context concept that integrates global context and deep context, which is mainly the connection of context. Therefore, it can capture the highquality long-distance contextual information between words and utterances better than global context and deep context:

\[ C = [C^1, \ldots, C^m] \]  

where \( C^m = [\bar{H}_1, \bar{H}_2, \ldots, \bar{H}_n] \). Then, using the context vector \( C \) to receive the \( \hat{Q} \) and \( \hat{K} \), and then obtain the output of the self-attention mechanism:

\[ \begin{bmatrix} \hat{Q} \\ \hat{K} \end{bmatrix} = \begin{pmatrix} 1 - \frac{\mu_Q}{\mu_K} \end{pmatrix} \begin{bmatrix} Q \\ K \end{bmatrix} + \frac{\mu_Q}{\mu_K} \begin{bmatrix} C \\ Z_Q \end{bmatrix} \]  

\[ \sqrt{\frac{\mu_Q}{\mu_K}} = \sigma \left( \begin{bmatrix} Q \\ K \end{bmatrix} \begin{bmatrix} F_Q^H \\ F_Q^K \end{bmatrix} + C \begin{bmatrix} Z_Q \\ Z_K \end{bmatrix} \begin{bmatrix} F_Q^C \\ F_K^C \end{bmatrix} \right) \]  

where \( \sigma(\cdot) \) is the logistic sigmoid function, \( F_Q^H, F_K^H, F_Q^K, F_K^K \) and \( F_Q^C, F_K^C \) are the trainable parameters with the size of \( d \times 1 \), the context vector \( C \) is \( n \times d_c \), \( Z_Q, Z_K \) are the associated trainable parameter with \( d_c \times d \). Where the Q, K, V can be obtained through the trainable parameter \( \{T_Q, T_K, T_V\} \):

\[ \begin{bmatrix} Q \\ K \\ V \end{bmatrix} = H \begin{bmatrix} T_Q \\ T_K \\ T_V \end{bmatrix} \]  

where the hidden layers are a shape of \( m \times d \), queries \( Q \), keys \( K \), and values \( V \) is a size of \( m \times d \), where \( T_Q, T_K, T_V \) are trainable parameter with a shape of \( d \times d \) and \( d \) is the scale of hidden layers. The O represents the output of self-attention mechanism:

\[ O = \text{ATT} \left( \hat{Q}, \hat{K} \right) V \]  

where \( \text{ATT}(\cdot) \) is a dot product attention model, which can save more space.
2.4. Connection Mechanism

The connection mechanism can connect the hidden layers of BiGRU with the word embedding and the utterance embedding, which can enhance the contextual embedding information of each word and utterance. Thus the contextual word embedding is modified to:

\[ H^c_j = \left[ H^j; e(w_j), H^j \right] \]  
\[ e_c(w_j) = \tanh(W_v \cdot H^c_j + l_u) \]

and the contextual utterance embedding is adjusted to:

\[ H^c_i = \left[ H^i; e(u_i), H^i \right] \]  
\[ e_c(u_i) = \tanh(W_q \cdot H^c_i + l_w) \]

3. Experiment

3.1. Datasets

There are two public datasets for evaluating proposed model, including Friends dataset [15], EmotionPush dataset [15]. The first dataset comes from the dialogues between people in the Friends TV script, and there are a total of 1000 dialogues, in which 720 dialogues are used for training, 80 dialogues are used to verify the model, and 200 dialogues are assigned to test. The second dataset is derived from the dialogues between friends on the Facebook Messenger. Among them, 720 dialogues are divided for training, 80 dialogues are used to verify the model, and 200 dialogues are assigned to test. Each utterance in these datasets has an emotional label, including joy, sadness, anger, surprise, disgust, neutral, non-neutral, and fear. we only evaluate the proposed model using anger, joy, sadness and neutral. The distribution of the four emotions in the Friends dataset and EmotionPush dataset is shown in Table 1.

3.2. Evaluation

For emotion recognition, we mainly apply the accuracy of four emotions and two evaluation metrics to evaluate the performance of the proposed model, namely weighted accuracy and unweighted accuracy:

\[ WA = \sum_{k=1}^{|k|} p_k \cdot a_k \]  
\[ UWA = \frac{1}{|k|} \sum_{k=1}^{|k|} a_k \]

where \( p_k \) is the proportion of the emotion class k in the testing dataset and \( a_k \) is the homologous accuracy of emotion class k.
Table 2. Performance of four emotion categories on the test dataset of Friends, the four emotion are anger, joy, sadness and neutral. The bold displays the best results

| Methods             | Anger(AC) | Joy(AC) | Sadness(AC) | Neutral(AC) | WA   | UWA   |
|---------------------|-----------|---------|-------------|-------------|------|-------|
| BcLSTM(10)          | 64.7      | 69.6    | 48.0        | 75.6        | 72.4(4.2) | 64.4(1.6) |
| SA-BiLSTM(11)       | 49.1      | 68.8    | 30.6        | **90.1**    | **79.8** | 59.6  |
| HiGRU(12)           | 62.8      | 69.6    | 52.1        | 70.1        | 68.6(9.7) | 63.7(3.9) |
| HiGRU-f(12)         | 63.5      | 70.5    | 53.6        | 73.3        | 71.1(2.0) | 65.2(2.5) |
| HiGRU-sf(12)        | 64.1      | 69.0    | 54.1        | 73.5        | 71.1(3.6) | 65.2(2.1) |
| PT-CoDE(13)         | 57.5      | 70.0    | 42.2        | 87.2        | 79.6(1.0) | 64.2(1.9) |
| BiGRU+global context| **64.9**  | **70.6**| **54.2**    | 72.0        | 70.5(3.4) | **65.4**(1.3) |
| BiGRU+deep context  | 62.4      | **71.6**| **54.7**    | 75.1        | 72.5(2.3) | **65.9**(1.8) |
| BiGRU+deep-global context | **65.6** | **72.3**| **58.2**    | 71.3        | 70.4(2.4) | **66.9**(1.5) |

Table 3. Performance of four emotion categories on the test dataset of EmotionPush, the four emotion are anger, joy, sadness and neutral. The bold displays the best results

| Methods             | Anger(AC) | Joy(AC) | Sadness(AC) | Neutral(AC) | WA   | UWA   |
|---------------------|-----------|---------|-------------|-------------|------|-------|
| BcLSTM(10)          | 32.9      | 69.9    | 47.1        | 78.0        | 74.7(4.4) | 57.0(2.1) |
| SA-BiLSTM(11)       | 24.3      | 70.5    | 31.0        | **94.2**    | **87.7** | 55.0  |
| HiGRU(12)           | 39.7      | 68.0    | 48.2        | 75.9        | 72.6(1.6) | 57.8(1.4) |
| HiGRU-f(12)         | 40.0      | 67.2    | 48.0        | 76.3        | 73.1(2.7) | 57.9(1.3) |
| HiGRU-sf(12)        | 40.3      | 66.7    | 48.6        | 77.7        | 74.0(2.6) | 58.3(1.2) |
| PT-CoDE(13)         | 16.8      | 64.0    | 31.6        | 91.4        | 83.0(0.8) | 50.9(1.0) |
| BiGRU+global context| **40.5**  | **70.6**| **48.8**    | 74.6        | 72.4(2.8) | **58.6**(1.7) |
| BiGRU+deep context  | 38.9      | 70.5    | **51.7**    | 73.9        | 72.0(2.4) | **58.8**(1.1) |
| BiGRU+deep-global context | **40.5** | **70.9**| **53.2**    | 71.4        | 70.2(3.2) | **59.0**(2.0) |

3.3. Baselines
For this task, we use six baselines, including BcLSTM [10], SA-BiLSTM [11], HiGRU [12], HiGRU-f [12], HiGRU-sf [12], and PT-CoDE [13]. The first method can extract textual features by using an ordinary convolutional layer and max-pooling. Then, the use of a bidirectional long short-term memory can generate the utterance sequence representations. The second model could learn utterance embedding, but the self-attention cannot catch the long-distance contextual information of words and utterances. The third model consisted of a lower GRU that produces word embeddings and an upper GRU that captures the contextual utterance embeddings in the dialogues. The fourth model fused the individual features, which can improve the overall performance of the model. The fifth model fused individual features fusion and self-attention, which is more efficient to learn contextual information than HiGRU, but the self-attention is unable to catch the high-quality long-distance contextual information of words and utterances. The last model pre-trained a context-dependent encoder for utterance-level emotion recognition by learning from unlabeled dialogue data.

3.4. Analysis
Compared with the state-of-the-art methods on the task, the results of the test dataset of Friends and EmotionPush are listed in Table 2 and Table 3, which use the round brackets to record the standard deviations of WA and UWA, the results of BcLSTM and SA-BiLSTM are from Poria et al [10] and Linkai et al [11]. From these tables, we can find that it’s obvious that the proposed model performs best compared with all baselines in utterance emotion recognition. As can be seen from Figure 3, the deep-global context effect is better than all methods. The BiGRU and the global context model have significant effects on the types of anger, joy
and sadness emotion. Although the performance on neutral emotion is lower than SA-BiLSTM, the UWA obtains 65.4 and 58.6 on the test dataset of Friends and EmotionPush, which is better than the baselines as well. The BiGRU and the deep context model performs better in joy and sadness emotion types than the baselines, and the result of UWA can reach 65.9 and 58.8 on the two datasets. Furthermore, the deep context model is better than the global context model in sadness emotion and UWA, which is improved by 0.5% and 2.9% of the sadness emotion and 0.5% and 0.2% of UWA. Even though the global context model and deep context model obtain the significant performance on anger, joy, and sadness emotion types and UWA, the BiGRU and the deep-global context model obtain better performance than above context model. The experimental results also reveal that the deep-global context model has the best performance on anger, joy and sadness, which has reached 65.6, 72.3, and 58.2 on accuracy for the test dataset of Friends and has reached 40.5, 70.9, 53.2 on accuracy for the test dataset of EmotionPush. Moreover, the UWA reached 66.9 and 59.0 on the test dataset of public Friends and EmotionPush. The results prove that the effectiveness of our proposed model, especially in the use of deep-global context model is more effective than other advanced baselines.

![Figure 3. Comparison of UWA on two datasets](image)

4. Conclusion
Emotion recognition is used extensively on dialogues, however, little is known about capturing the high-quality long-distance contextual information in dialogues. The bidirectional gated recurrent units generate contextual word embedding and contextual utterance embedding. The gated recurrent unit network has also made good achievements in pattern recognition, for example, the use of convolutional neural networks and gated recurrent unit networks can study the travel characteristics and laws of residents. We used three contexts in the self-attention mechanism, namely global, deep and deep-global context, which can enhance the high-quality long-distance contextual information of words and utterances by capturing the overall meaning of the words and utterances. We conducted emotion recognition task for evaluating the proposed approach and the experimental results show that the deep-global context model is more effective compared with the state-of-the-art baselines. As future work, we plan to use the BERT model to train for the utterances and add video and audio features for multimodal features emotion recognition tasks.
Acknowledgments
This research is supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN201900620) and the Educational Reform Project of Chongqing Municipal Education Commission (Grant No. 203780).

References
[1] De Choudhury M., Gamon M., Counts S., et al. Predicting depression via social media[C]//Proceedings of the International AAAI Conference on Web and Social Media. Boston, USA: Association for the Advancement of Artificial Intelligence, 2013: 1-10.
[2] Cambria E., Poria S., Gelbukh A., et al. Sentiment analysis is a big suitcase[J]. IEEE Intelligent Systems, 2017, 32(6): 74-80.
[3] Xing F. Z., Cambria E., Welsch R. E. Natural language based financial forecasting: a survey[J]. Artificial Intelligence Review, 2018, 50(1): 49-73.
[4] Young T., Cambria E., Chaturvedi I., et al. Augmenting end-to-end dialog systems with commonsense knowledge[J]. 2017, arXiv: 1709.05453.
[5] Ekman P, Friesen W V. Constants across cultures in the face and emotion[J]. Journal of personality and social psychology, 1971, 17(2): 124.
[6] Ekman, Paul. Strong evidence for universals in facial expressions. A reply to Russell’s mistaken critique[J]. Psychological Bulletin, 1994, 115(2): 268-287.
[7] Olson D. From utterance to text: The bias of language in speech and writing[J]. Harvard educational review, 1977, 47(3): 257-281.
[8] Golder S A, Macy M W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures[J]. Science, 2011, 333(6051): 1878-1881.
[9] Rozgić V., Ananthakrishnan S., Saleem S., et al. Ensemble of svm trees for multimodal emotion recognition[C]//Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. New York, USA: Institute of Electrical and Electronics Engineers, 2012: 1-4.
[10] Poria S., Cambria E., Hazarika D., et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 873-883.
[11] Linkai Luo, Haiqing Yang, and Francis Y. L. Chin. Emotionx-dlc: Self-attentive BiLSTM for detecting sequential emotions in dialogues[C]//Natural Language Processing for Social Media. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 32-36.
[12] Jiao W., Yang H., King I., Lyu MR (2019) HiGRU: hierarchical gated recurrent units for utterance-level emotion recognition[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 397–406.
[13] Jiao W., Lyu M. R., King I. PT-CoDE: Pre-trained Context-Dependent Encoder for Utterance-level Emotion Recognition[J]. 2019, arXiv: 1910.08916v1.
[14] Yang B, Li J, Wong D F, et al. Context-aware self-attention networks[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence. Hawaii, USA: Association for the Advancement of Artificial Intelligence, 2019: 387-394.
[15] Hsu CC, Chen SY, Kuo CC, Huang TH, Ku LW. EmotionLines: an emotion corpus of multi-party conversations[C]//Proceedings of the 11th International Conference on Language Resources and Evaluation. New Orleans, USA: European Language Resources Association, 2018: 1597-1601.