Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

DHF-Net: A hierarchical feature interactive fusion network for dialogue emotion recognition

Chenquan Gan [a,b], Yucheng Yang [b], Qingyi Zhu [a], Deepak Kumar Jain [c], Vitomir Struc [d,*]

[a] *School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
[b] *School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
[c] *Key Laboratory of Intelligent Air Ground Cooperative Control for Universities in Chongqing, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China, Chongqing 400065, China*
[d] *Faculty of Electrical Engineering, University of Ljubljana, Trzaska cesta 25, SI-1000, Ljubljana, Slovenia*

ARTICLE INFO

Keywords:
Dialogue emotion recognition
Contextual information
Fine-grained information
Hierarchical feature
Interactive fusion

ABSTRACT

To balance the trade-off between contextual information and fine-grained information in identifying specific emotions during a dialogue and combine the interaction of hierarchical feature related information, this paper proposes a hierarchical feature interactive fusion network (named DHF-Net), which not only can retain the integrity of the context sequence information but also can extract more fine-grained information. To obtain a deep semantic information, DHF-Net processes the task of recognizing dialogue emotion and dialogue act/intent separately, and then learns the cross-impact of two tasks through collaborative attention. Also, a bidirectional gate recurrent unit (Bi-GRU) connected hybrid convolutional neural network (CNN) group method is designed, by which the sequence information is smoothly sent to the multi-level local information layers for feature exaction. Experimental results show that, on two open session datasets, the performance of DHF-Net is improved by 1.8% and 1.2%, respectively.

1. Introduction

With the rapid development of intelligent human–computer interaction systems, the task of dialogue emotion recognition is becoming increasingly crucial for downstream tasks, such as emotional perception chat agent (Chen et al., 2018; Zhou et al., 2020), persuasive dialogue (Chen et al., 2021), visual question and answer (Tapaswi et al., 2016; Ye et al., 2021), and health dialogue (Althoff et al., 2016; Olander & Koinberg, 2017). This task needs to understand the semantics of the dialogue and recognize the emotional changes of the speaker.

In the past, many methods sought the relationship between words in discourse from the word level and extracted features to obtain an emotional classification. The commonly used technologies include synonym networks and emotion dictionaries. After the widely use of machine learning, other methods have been applied, such as the improved clustering algorithm (Borlea et al., 2021), rule-based statistical method (Chiang et al., 2014), naive Bayes model (Alangari & Alturki, 2020), and support vector machine combined with other domain methods (Upadhyay & Nagpal, 2020).

With the widespread application of deep learning techniques, researchers try to extract features from contextual information through neural network models. Contextual information can be extracted from the local historical discourse of adjacent sentences or the remote discourse history of the whole conversation. In addition to contextual information, the factors affecting conversational emotion recognition also include the act/intent form speaker, which has also been confirmed by Kumar (Kumar et al., 2018). Considering these factors, recently, these models are often based on artificial neural networks combined with manual features (Albu et al., 2019; Tan et al., 2014), or using convolutional neural networks, recurrent neural networks, attention mechanisms, and even graph convolutional networks (Poria, Majumder, Mihalcea & Hovy, 2019).

Unfortunately, in existing methods, the understanding of deep semantic information in the dialogue is insufficient, and there is a conflict in extracting features at the context level and fine-grained level. To solve these problems, this paper proposes a hierarchical feature interactive fusion network (defined DHF-Net). Through the GRU module and the designed multi-level flat convolution group, DHF-Net respectively extracts emotion features and act/intent features at the context level and fine-grained level, then uses the collaborative attention mechanism to integrate these two features interactively, so as to finally output emotion classification. Experimental results show its superiority.

[*] Corresponding author.
E-mail addresses: gcq2010cqu@163.com (C. Gan), s190101138@stu.cqupt.edu.cn (Y. Yang), zhuqy@cqupt.edu.cn (Q. Zhu), deepak@cqupt.edu.cn (D.K. Jain), vitomir.struc@fe.uni-lj.si (V. Struc).

https://doi.org/10.1016/j.eswa.2022.118525
Received 7 December 2021; Received in revised form 31 July 2022; Accepted 10 August 2022
Available online 16 August 2022
0957-4174/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

In summary, our contribution mainly includes the following three parts:

(1) The proposed method DHF-Net conducts a comprehensive analysis from the two aspects of emotion and behavioral intention, which is different from the previous method that only starts from emotion modeling and calculates the intermediate variables that affect emotion.
(2) We design a feature extraction method in which the hybrid CNN group connects the Bi-GRU, which smooths the fine-grained hierarchical features while retaining the context-level features and alleviates the conflict between them.
(3) We obtain deep semantic information by interactively fusing the extracted emotional and behavioral intent features through a collaborative attention mechanism.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes the proposed method in detail. Section 4 exhibits some experiments to verify the proposed method and the analysis of experimental results. Finally, Section 5 summarizes this paper.

2. Related work

Early studies extracted features at the word level to analyze sentences. The typical methods are synonym networks (Strapparava & Valitutti, 2004), emotion dictionaries (Mohammad, 2018; Wu et al., 2018), subsequent word embedding technology (Mikolov et al., 2013; Pennington et al., 2014). However, these processing methods lack the understanding of the context in continuous dialogue. After the popularity of deep learning, the methods of extracting emotion features at the context level and fine-grained level are the two main directions.

Extracting features at context level is the most common deep learning method in dialogue emotion analysis, which makes use of the advantages of recurrent neural network (RNN) in processing sequence information (Young et al., 2020). In Hazarika, Poria, Zadeh, et al. (2018), an external memory module was introduced in the conversation memory network to enhance the extraction of conversation history information. In Poria et al. (2017), the bidirectional long short-term memory (Bi-LSTM) was designed to process the target text as well as the above and infra utterances, which aims at context modeling to capture the influence of context. In Hazarika, Poria, Mihalcea, Cambria and Zimmermann (2018), the gate recurrent unit (GRU) based on a memory network was added to process each sentence between two interlocutors. In Jiao et al. (2019), the hierarchical GRU was used to model word-level semantics and sentence-level context. The work (Majumder et al., 2019) proposed DialogueRNN, which tracks the state and emotion changes of each speaker by multiple specific GRUs and contextual information. These methods focus on contextual information and speaker state modeling, but they fail to obtain deep semantics at a fine-grained level.

Extracting features at a fine-grained level is also widely used, which capitalizes convolutional neural network (CNN) to obtain deep semantic advantages of discourse itself. In Kim (2014), CNN was first utilized to extract features at the fine-grained level for emotion classification. In Bertero et al. (2016), the author applied CNN to infer emotion from the current conversation without contextual sentences. In DialogueGCN (Ghosal et al., 2019), the author used a graph convolution neural network (GCN) to capture emotional jumps in short conversations. These methods are very sensitive to emotional changes in continuous dialogue and achieve good results, but they are weak in considering the influence of context, which makes it difficult to model the effect of long-distance semantic emotion.

Noting the limitations of the above-mentioned methods and being inspired by the fact that dialogue act/intent recognition plays a significant role in understanding the meaning of dialogue and judging its emotion (Blache et al., 2020; Cerisara et al., 2018; Qin et al., 2020), this paper proposes a hierarchical feature interactive fusion network (named DHF-Net). DHF-Net aims at a trade-off between extracting contextual and fine-grained features, and interactively fusing dialogue sentiment and behavior/intent information for deep semantics. We will describe DHF-Net in detail in the next section.

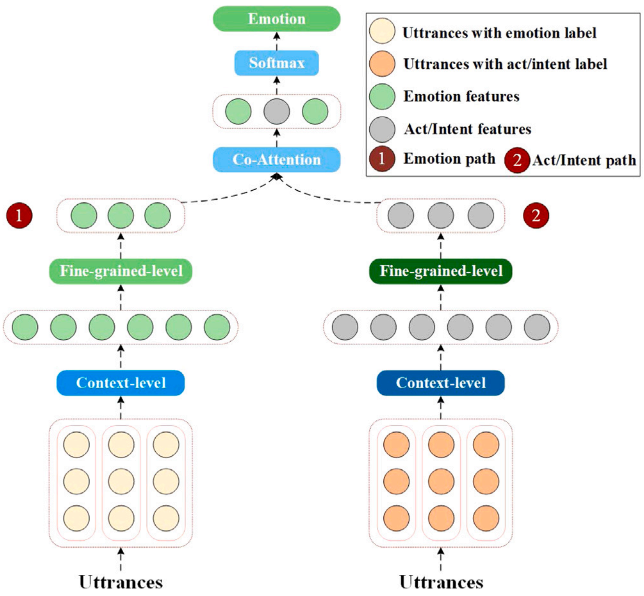

Fig. 1. The framework of DHF-Net. The left and right are the extraction paths for emotional features and act/intent features, respectively.

3. The proposed DHF-Net

The framework of DHF-Net is shown in Fig. 1. The left side is the extraction path of emotion features, and the right side is the extraction path of act/intent features. The modules and their functions are as follows. First, the input module embeds the word vector form utterance into the input model. Second, the context level feature extraction module extracts features at the context level using Bi-GRU. Third, the fine-grained level extraction module extracts features at the fine-grained level using a hybrid CNN group. Next, the hierarchical feature interactive fusion module interactively fuses the features on two sides through the collaborative attention mechanism. Finally, the classification module outputs the classification label.

3.1. Context level feature extraction

The context-level feature extraction plays a crucial role in the dialogue because long-distance utterances in continuous dialogue have a clear impact on current emotion. So here select the appropriate RNN as the appropriate context-level processing module. To match the preprocessing format and reduce the amount of calculation, Bi-GRU with fewer calculation parameters is adopted. It is the same Bi-GRU structure for emotion features and act/intent features extraction at context-level feature extraction, and only a few training parameters are set differently. The whole process of context-level feature extraction is shown in Fig. 2.

Each word in a fixed-length discourse is replaced by a word vector and embedded into the model through a short preprocessing. Here, Glove word embedding is used. Words not included in the pre-training word vector will be replaced with a random word vector, and the stop character or sentence end punctuation is reset to a fixed word vector for model training. The embedding process is shown as:

$$X_i = \text{Embedding}(u_i), \tag{1}$$

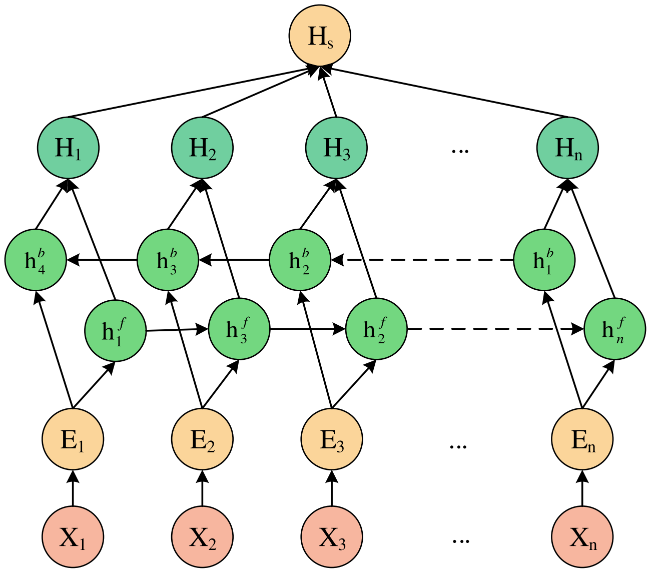

Fig. 2. Structure of context level feature extraction.

where $u_i = \{u_1, u_2, \dots, u_n\}$ is the ith input sentence and $X_i = \{X_1, X_2, \dots, X_n\}$ is the word embedding vector.

The word vector X_i is input into Bi-GRU including forward and backward GRU layers to extract the contextual information. The GRU layer includes an update gate z_i, a reset gate r_i, and an alternative gate $\tilde{h}_i$. Through gating control, the selective flow of information is realized. The internal information update of GRU layer is as:

$$z_i = \sigma(w^z x_i + u^z h_{i-1} + b^z), \tag{2}$$

$$r_i = \sigma(w^r x_i + u^r h_{i-1} + b^r), \tag{3}$$

$$\tilde{h}_i = \tanh(w^h x_i + r_i \odot u^h h_{i-1} + b^h), \tag{4}$$

$$h_i = z_i \odot \tilde{h}_i + (1 - z_i) \odot h_{i-1}, \tag{5}$$

where x_i is the input, h_{i-1} is the hidden state of GRU layer, $\sigma(\cdot)$ is the activation function, and the symbol $\odot$ denotes the element-wise multiplication. Among these equations, the w^z, u^z and b^z are the input weight, hidden layer weight and offset in the update gate, respectively. The role of w^r, u^r, b^r in reset gate and the role of w^h, u^h, b^h in candidate state are consistent with w^z, u^z and b^z. The output of Bi-GRU hidden state is as:

$$H_t = h_t^f + h_t^b, \tag{6}$$

$$H_s = \mathrm{add}(H_1, H_2, \dots, H_i), i = 1, 2, \dots, t, \tag{7}$$

where h_t^f and h_t^b represent the forward GRU and the backward GRU, respectively. H_t is the output of each GRU, the summation result H_s is sent to the fine-grained level feature extraction module for further processing.

3.2. Fine-grained level feature extraction

For both emotion and act/intent features, it is not enough only through the context level feature extraction, the fine-grained level feature extraction is also necessary. Using the conventional CNN method will destroy the existing sequence information because it is over-extracted at the fine-grained level, which will damage the feature coming from the context level. In this paper, a series of smooth hybrid CNN groups are utilized to connect the context level features and fine-grained level features, which can reduce the conflict between them. The whole extraction process of the fine-grained level is displayed in Fig. 3.

The hybrid CNN groups consist of Conv1D, hyperbolic linear units (HLU) activation function, and average pooling. In the feature extraction process, the effect of one-dimensional convolution is more gentle than that of high-dimensional convolution, which leads to less damage to sequence information.

$$H_c = \mathrm{Conv1D}(H_s), \tag{8}$$

$$H_p = \mathrm{average-pooling}(H_c), \tag{9}$$

$$H_f = \mathrm{HLU}(H_p), \tag{10}$$

where H_c is the output of convolution layer, H_p is the output of pooling layer, H_f is the output after HLU activation function.

The HLU function (Li et al., 2016) is selected for activation, and the average output of HLU activation function is close to 0, which can reduce the offset between natural gradient and normal gradient and make the convolution process smoother. The equation of HLU function is as:

$$f(x) = \begin{cases} x, & x > 0, \\ \alpha x/(1-x), & x < 0, \end{cases} \tag{11}$$

where x is the value passed by the upper layer and α is the adjustable parameter.

Compared with the maximum pool method, the average pool method can expand the range of perception, and retain fine-grained level features and long-distance context level features with less information loss. In the emotion feature extraction part, fine-grained features are processed in series by three hybrid CNN groups, and then the output is sent to the interaction fusion part.

Considering that emotion is contained in the deep semantics of utterances and act/intent belongs to a more intuitive expression, here select three hybrid CNN groups for thoroughly processing in the emotion feature extraction part, and reduce the number to two-hybrid CNN groups in act/intent feature extraction part to meet the demand.

3.3. Hierarchical feature interactive fusion

The purpose of hierarchical feature interactive fusion is to enrich deep semantic information by making comprehensive use of emotion features and act/intent features in dialogue. To this end, here choose the collaborative attention (Co-attention) mechanism (Li et al., 2018). Different input features exchange partial features after multiple training, that is, they carry both information at the same time. After the subsequent fully connected layer and classification layer processing, all information is comprehensively used to achieve the effect of hierarchical feature interactive fusion and more accurate recognition of dialogue emotion. The whole hierarchical feature interactive fusion process is demonstrated in Fig. 4. The output of the interactive fusion part is as:

$$E_h = \mathrm{softmax}(H_e), \tag{12}$$

$$A_h = \mathrm{softmax}(H_a), \tag{13}$$

$$E_{h'} = \mathrm{reshape}(E_h), \tag{14}$$

$$C_q = E_h A_h, \tag{15}$$

$$C_h = C_q E_{h'}, \tag{16}$$

$$I_f = \mathrm{concat}(\mathrm{E}_{h'}, \mathrm{C}_h), \tag{17}$$

where H_e and H_a are output by emotion branch and act/intent branch, respectively. I_f is the fusion features, $E_{h'}$ is the compressed emotional feature, C_q and C_h are the hidden layer features after point multiplication in the Co-attention part. Emotion features E_h and act/intent features A_h through two product processes obtain C_q and C_h, and then concatenate with compressed $\mathrm{E}_{h'}$. The fused features contain two types of information.

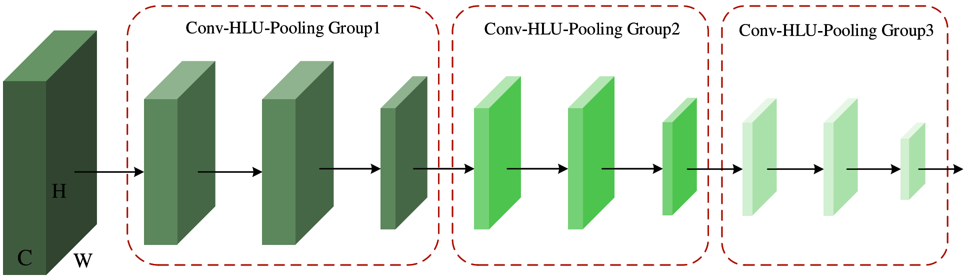

Fig. 3. Structure of fine-grained level feature extraction based on the Conv-HLU-Pooling method, C represents the number of channels, H and W represent the data dimensions. Multiple convolution groups are used in series, and each group has the same structure while the parameters need to be adjusted as necessary.

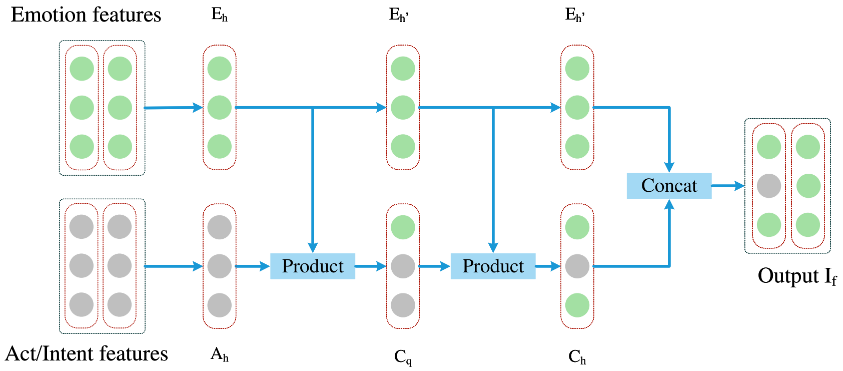

Fig. 4. The interactive fusion process of hierarchical features in the Co-attention mechanism.

The fused features have output after fully connected layer, and the weights are redistributed through the self-attention layer. According to the standard attention equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{dk}}\right)V, \tag{18}$$

where Q, K and V are obtained from the output of the previous hidden layers, respectively. Therefore, the output of a fully connected layer after weight allocation is as:

$$y_i = \tanh(w^y y_t + b^y), \tag{19}$$

$$\text{y}_t = \text{Attention}(Q, K, V), \tag{20}$$

where w^y and b^y are trainable weights and offsets, respectively. Though the connection result, the softmax classifier outputs the probability of each emotion and takes the highest probability as the final output emotion:

$$y_o = \text{softmax}(\text{y}_i). \tag{21}$$

The cross-entropy is selected for the loss function. The function expression is as:

$$L = -[y\log_2 \hat{y} + (1-y)\log_2(1-y)], \tag{22}$$

where y is the output value of the model, $\hat{y}$ is the tag value of the test set.

4. Experiments

4.1. Datasets and evaluation metrics

To verify the effectiveness of DHF-Net, here benchmark the network based on a large-scale public conversation dataset Dailydialog (Li et al., 2017) with emotion and act/intent tags and MELD (Multimodal EmotionLines Dataset) (Poria, Hazarika, Majumder & Naik, 2019) dataset with emotion tags. These two datasets are typical datasets in emotion recognition in dialogue tasks, the distribution of discourse and tags is shown in Table 1.

Table 1
Comparison between Dailydialog and MELD datasets.

Dataset	Dailydialog	MELD
Dialog	13,118	1,433
Utterance	103,607	13,708
Training	11,118	9,989
Validation	1,000	1,039
Test	1,000	2,610

Dailydialog dataset[1] covers all kinds of topics in our daily life, including health, emotion, education, life, and so on. Emotions can be divided into seven labels: fear, disgust, anger, neutral, happiness, sadness, and surprise. Act tags can be divided into four categories: statement, inquiry, instruction, and commitment.

MELD dataset[2] comes from 13000 sentences in 1433 dialogues of the TV series friends. Each sentence corresponds to an emotion label, which is divided into seven categories: neutral, angry, disgusted, joy, fear, sadness, and surprise. MELD contains very short single utterances, such as yeah, oh.

For evaluation metrics, Accuracy and Macro-F1 score are chosen. The calculation formula is as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{23}$$

$$\text{F1}_i = \frac{2TP}{2TP + FN + FP}, \tag{24}$$

$$\text{Macro-F1} = \frac{1}{n}\sum_{i=1}^{n}\text{F1}_i, \tag{25}$$

where TP, FP, FN, and TN represent the number of true positive, false positive, false negative, true negative, positive, and negative, respectively. F1_i represents F1 for each label.

[1] http://yanran.li/dailydialog.

[2] http://affective-meld.github.io.

Table 2
The important parameters used in training.

Parameter	Value
Embedding size	300
Hidden size	256
Batch size	64
Dropout	0.2
Learning rate	10^{-6}
Optimization	Adam

Accuracy is used to measure the classification accuracy of all tags, Macro-F1 is used to measure the effectiveness of each category label, which is suitable for the situation where label distribution is not completely balanced. Take the average of the results obtained after more than 10 test sets as the final result.

4.2. Training setup and comparison models

Our model is trained on NVIDIA Tesla V100 32 GB GPU. Keras 2.2.4 and TensorFlow 1.50 are selected for the deep learning framework. The basic parameters settings are shown in Table 2. Training, validation, and test sets are used directly without re-partitioning because the dataset has been divided at the time of publication. After multiple validations on the test set, the arithmetic mean of the results is taken as the final result.[3]

To reduce the diversity of different length utterances during training, we align all utterances according to the maximum length and fill in the insufficient content with 0. The input utterance is embedded through the Glove model, where the dimension $d = 300$. The hidden layer size is set to 256 to match the word embedding dimension. The dimensions of feature extraction layers are also matched with the previous layer and reduced gradually. A small learning rate and the corresponding batch size are selected to make the model fully trained (see Table 2).

To prevent performance degradation caused by overfitting, the dropout rate is set to 0.2, and the early stopping patience value is to 10 so that the training will stop when the performance of the model is no longer improved within 10 rounds. The Adam (adaptive motion estimation) optimization algorithm (Chilimbi et al., 2014) is adopted, the learning rate is dynamically adjusted according to the moment estimation of the gradient during training, and the gradient oscillation is reduced in the samples with uneven label distribution, so as to obtain a good training performance.

To demonstrate the effectiveness of our method, several typical baseline methods in the dialogue emotion recognition task are selected for comparison:

- **Glove CNN** (Kim, 2014). A model based on word embedding and CNN.
- **IIIM** (Kim & Kim, 2018). The comprehensive neural network model based on CNN, which is skilled in extracting fine-grained semantics in dialogue.
- **BC-LSTM** (Poria et al., 2017). Based on the improved bidirectional LSTM model, which does well in extracting features at context level.
- **ICON** (Hazarika, Poria, Mihalcea, Cambria & Zimmermann, 2018). Combining RNN and memory neural network, which performs well in context feature extraction.
- **DialogueRNN** (Majumder et al., 2019). The hierarchical RNN model of multiple GRUs, which can be used to model context information and speaker state changes.
- **JointDAS** (Cerisara et al., 2018). A hierarchical recursive network, which can perform dual task processing of emotion and dialogue behavior recognition.
- **DCR-Net** (Qin et al., 2020). Based on the deep relational network, excellent performance in joint modeling of emotion and dialogue behavior tasks.
- **Bif-BiAGRU** (Jiao et al., 2020). Combining the two-way GRU and the improved attention weight update model, and is good at long-distance context processing.
- **KET** (Zhong et al., 2019). The method based on transformer structure which enhances the effect of emotion recognition with the help of context information and external common sense knowledge.

4.3. Experimental results and analysis

Tables 3 and 4 show the comparison results with the above-mentioned models on Dailydialog dataset and MELD dataset. Table 5 gives the complexity analysis. Figs. 5–7 display the label classification confusion matrix of our method. Now, let us introduce these results and analysis in detail.

Table 3 demonstrates the comparison results of DHF-Net and the other six models in two tasks on the Dailydialog dataset. The results exhibit that the Accuracy and Macro-F1 of DHF-Net on the emotion (act) label are 1.8% and 2.9% (0.2% and 2.7%) better than the other best results, respectively. This is because that DHF-Net not only has the task interaction of act/intent and emotion recognition but also improves the effect of feature extraction at the context level and fine-grained level. Glove CNN (Kim, 2014) and IIIM (Kim & Kim, 2018) only improved the feature extraction method at the fine-grained level, JointDAS (Cerisara et al., 2018) regarded the act/intent and emotion recognition as independent tasks without further exploration, BC-LSTM (Poria et al., 2017) and DialogueRNN (Majumder et al., 2019) did not interact act/intent features with emotional features, DCR-Net (Qin et al., 2020) lacked the improvement of fine-grained level and context level extraction, so their performance is inferior to DHF-Net.

Table 4 lists the comparison results of DHF-Net and five models on the MELD dataset. There is no explicit act/intent label in the MELD dataset, so five models with good performance in emotion recognition tasks are selected in this experiment. The results display that the Accuracy and Macro-F1 of DHF-Net on the emotion label are 1.2% and 1.5% better than the other best results, respectively. This can be explained that, although there is no explicit act/intent label in the MELD dataset, the trained DHF-Net can still extract act/intent features and improve the effect of emotion recognition through interaction, and finally obtain the best performance. ICON (Hazarika, Poria, Mihalcea, Cambria & Zimmermann, 2018) only improved feature extraction at the context level. The Bif-BiAGRU (Jiao et al., 2020) which used the HMN module had not captured deep semantics. BC-LSTM (Poria et al., 2017) and Glove DialogueRNN (Majumder et al., 2019) focused on speaker state changes without the interaction of act/intent. KET (Zhong et al., 2019) used transformer framework to improve the model, but still did not pay attention to speaker intention semantics, so they gain lower performance than DHF-Net. For DHF-Net, the effect of interactive fusion of two tasks is also evidently better than that of a single task.

Table 5 illustrates the complexity comparison results. Because most methods in Tables 3 and 4 lack detailed training settings, it is hard to obtain their complexity data. Therefore, three types of typical deep learning structure methods are selected, including Text-CNN (Guo et al., 2019) (a vanilla CNN framework), DialogueRNN (Majumder et al., 2019) (a baseline model of RNN framework), and Ntuer (Zhong & Miao, 2019) (a common large-scale CNN framework). Besides, the parameters and FLOPs (Floating-point operations per second) (Molchanov et al., 2017) are both considered, which measure the computational power required by a model.

As can be seen from Table 5, the complexity of DHF-Net is between the RNN model and the large-scale 3D CNN model. Compared with the baseline model DialogueRNN, DHF-Net has fewer parameters, fewer computing power requirements, and better classification performance. This means that it is better and easier to deploy on other small devices.

[3] The model and training related codes in https://github.com/yycpasserby/DERexperiments.

Table 3
Comparison results on Dailydialog dataset.

Method	Emotion		Act/Intent	
	Accuracy (%)	Macro-F1 (%)	Accuracy (%)	Macro-F1 (%)
JointDAS (Cerisara et al., 2018)	35.4	31.2	76.2	75.1
IIIM (Kim & Kim, 2018)	38.9	33	76.5	75.7
Glove CNN (Kim, 2014)	37.83	36.87	76.39	72.07
BC-LSTM (Poria et al., 2017)	40.35	39.27	81.87	79.12
DialogueRNN (Majumder et al., 2019)	44.5	41.8	82.15	79.6
DCR-Net (Qin et al., 2020)	47.3	45.4	80.2	79.1
DHF-Net (Ours)	**49.13**	**48.29**	**82.3**	**82.36**

Table 4
Comparison results on MELD dataset.

Method	Accuracy (%)	Macro-F1 (%)
BC-LSTM (Poria et al., 2017)	57.5	55.4
ICON (Hazarika, Poria, Mihalcea, Cambria & Zimmermann, 2018)	56.1	54.6
DialougeRNN (Majumder et al., 2019)	56.1	55.9
Bif-BiAGRU (Jiao et al., 2020)	59.2	57.1
KET (Zhong et al., 2019)	60.8	58.1
DHF-Net (Single)	**59.4**	**57.6**
DHF-Net (Ours)	**62.0**	**59.6**

Table 5
Complexity comparison.

Method	Params (10^5)	FLOPs (10^6)
Text-CNN (Guo et al., 2019)	1.51	9
DialogueRNN (Majumder et al., 2019)	30.1	216
Ntuer (Zhong & Miao, 2019)	84	479
DHF-Net (Single)	**14.1**	**60.4**
DHF-Net (Ours)	**28.4**	**141.4**

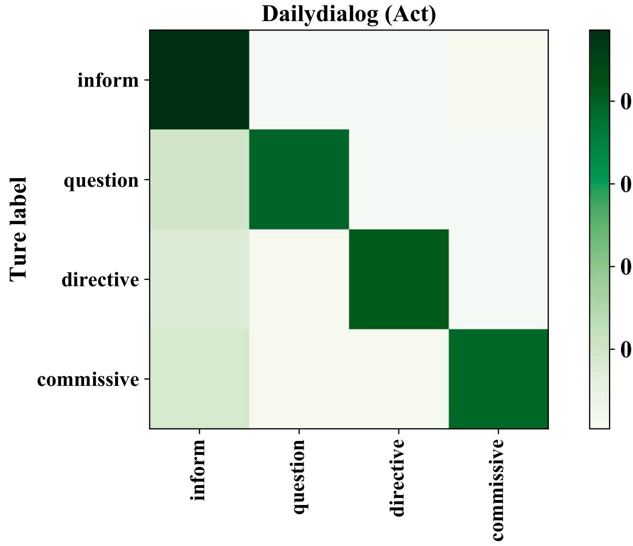

Fig. 5. Confusion matrix of act label classification in Dailydialog dataset.

The vanilla CNN framework has low complexity but poor performance. The large-scale 3D CNN framework has huge complexity overhead but may not break through the optimal performance. The RNN model is a better choice when considering both performance and complexity in dialogue emotion recognition in dialogue, but DHF-Net is optimized on the RNN model.

Figs. 5 and 6 show the label classification confusion matrix of two categories in the Dailydialog dataset. Among four tags of the act category, the inform label has the best classification performance, and the accuracy of the other three labels, i.e., question, directive, and commissive are around 0.8. Among seven emotion labels, neutral is the best, followed by anger, disgust, and happiness. Fear, sadness, and surprise get low classification accuracy. This implies that the distribution of act/intent categories is more uniform, and the shallow semantic information is easy to extract, so the overall classification performance is better. For different emotion categories, when the utterances are sufficient, the classification performance of DHF-Net is better and vice versa. The neutral label performs best, which is also consistent with the small proportion of emotional fluctuating discourse in the real life. Among the tags with poor classification performance, the fear label accounts for the smallest proportion of the dataset, and insufficient training leads to misjudgment. The surprise label is semantically similar to happiness and is easy to be misjudged. The sadness label in the Dailydialog dataset is more inclined to a slow emotional change, which is more difficult to identify than anger.

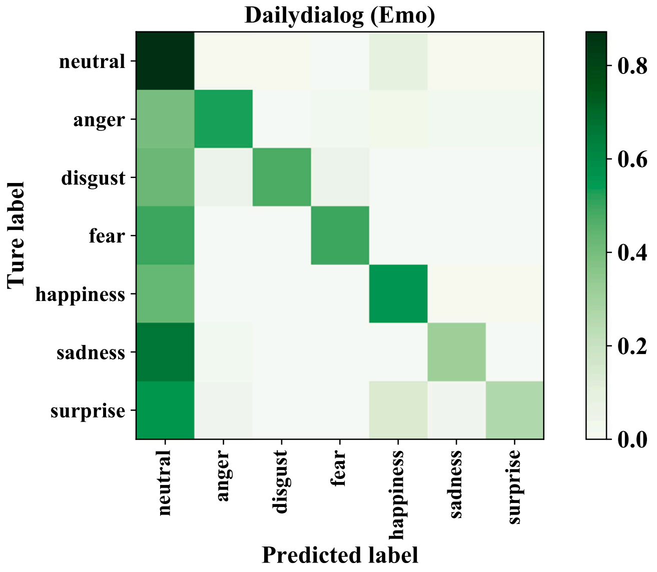

Fig. 6. Confusion matrix of emotion label classification in Dailydialog dataset.

Fig. 7 exhibits the confusion matrix of emotion label classification in the MELD dataset. It can be seen that Neutral, anger, happiness, sadness, and surprise have high classification accuracy, and the results of these five categories are similar. Disgust and fear get low classification accuracy. This implies that DHF-Net obtains good recognition performance on another multi-round dialogue dataset, and proves its generalization ability. Due to the different scenes of dialogue corpus sources, the classification performance of each label on the MELD dataset is different from that on the Dailydialog dataset. MELD dataset contains TV series dialogue discourse, leading to amplifying the explicit expression of some emotions such as sadness, much easier to identify than daily conversation. For the labels of a small proportion of samples, especially disgust and fear, there are only 271 and 268 of 9989 training labels, resulting in weak training performance and low classification accuracy than that of the labels in large proportion. In general, due

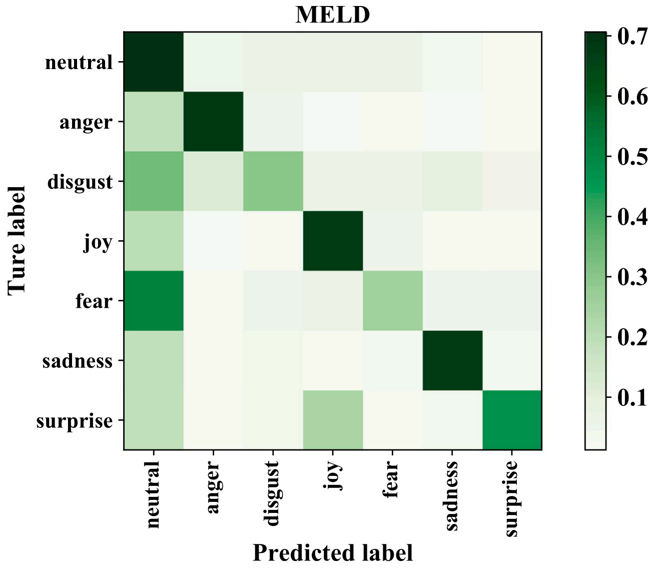

Fig. 7. Confusion matrix of emotion label classification in MELD dataset.

to the small number of training samples, emotion recognition of rare emotions is more difficult than ordinary emotions.

5. Conclusions

In this paper, a new network DHF-Net has been proposed, which is applied to the dialogue text based on multiple rounds of dialogue paragraphs for emotion recognition. After interactive processing of emotion features and act/intent features, it can get deep semantics and realize emotion classification. DHF-Net utilizes Bi-GRU and CNN-HLU-Pooling groups to extract features both at the context level and fine-grained level, then interactively fuse through the collaborative attention mechanism. Through the analysis and discussion of experimental results, it is confirmed that the joint modeling and interactive processing of emotion classification and act/intent judgment in dialogue can improve the final performance.

CRediT authorship contribution statement

Chenquan Gan: Writing – original draft, Data curation, Methodology, Resources, Software. **Yucheng Yang:** Writing – original draft, Formal analysis, Investigation, Methodology, Data curation. **Qingyi Zhu:** Writing – review & editing, Conceptualization, Methodology, Visualization, Investigation. **Deepak Kumar Jain:** Resources, Writing – review & editing, Supervision. **Vitomir Struc:** Writing – review & editing, Formal analysis, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Publicly available datasets were used.

Acknowledgments

The authors are grateful to the anonymous reviewers and the editor for their valuable comments and suggestions. This work is supported by Natural Science Foundation of China (Grant No. 61903056), Major Project of Science and Technology Research Program of Chongqing Education Commission of China (Grant No. KJZDM201900601), Chongqing Research Program of Basic Research and Frontier Technology, China (Grant Nos. cstc2019jcyj-msxmX0681, cstc2021jcyj-msxmX0761, and cstc2021jcyj-msxmX0530), Project Supported by Chongqing Municipal Key Laboratory of Institutions of Higher Education (Grant No. cqupt-mct-201901), Project Supported by Chongqing Key Laboratory of Mobile Communications Technology (Grant No. cqupt-mct-202002), Project Supported by Engineering Research Center of Mobile Communications, Ministry of Education (Grant No. cqupt-mct202006), and the ARRS Research Program P0250-B Metrology and Biometric Systems.

References

Alangari, N., & Alturki, R. (2020). Predicting students final GPA using 15 classification algorithms. *Romanian Journal of Information Science and Technology*, *23*, 238–249.

Albu, A., Precup, R. E., & Teban, T. A. (2019). Results and challenges of artificial neural networks used for decision-making and control in medical applications. *Facta Universitatis. Series: Mechanical Engineering*, *17*, 285–308.

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, *4*, 463–476.

Bertero, D., Siddique, F. B., Wu, C. S., Wan, Y., Ho, R., Chan, Y., & Fung, P. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1042–1047). EMNLP.

Blache, P., Abderrahmane, M., Rauzy, S., Ochs, M., & Oufaida, H. (2020). Two-level classification for dialogue act recognition in task-oriented dialogues. In *Proceedings of the 28th international conference on computational linguistics* (pp. 4915–4925). ICCL.

Borlea, I. D., Precup, R. E., Borlea, A. B., & Iercan, D. (2021). A unified form of fuzzy C-means and K-means algorithms and its partitional implementation. *Knowledge-Based Systems*, *214*, 106731–106740.

Cerisara, C., Jafaritazehjani, S., Oluokun, A., & Hoa, L. (2018). Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th international conference on computational linguistics* (pp. 745–754). COLING.

Chen, H., Ghosal, D., Majumder, N., Hussain, A., & Poria, S. (2021). Persuasive dialogue understanding: The baselines and negative results. *Neurocomputing*, *431*, 47–56.

Chen, C. H., Lee, W. P., & Huang, J. Y. (2018). Tracking and recognizing emotions in short text messages from online chatting services. *Information Processing & Management*, *54*, 1325–1344.

Chiang, H. S., Shih, D. H., Lin, B., & Shih, M. H. (2014). An APN model for arrhythmic beat classification. *Bioinformatics*, *30*, 1739–1746.

Chilimbi, T., Suzue, Y., Apacible, J., & Kalyanaraman, K. (2014). Project adam: Building an efficient and scalable deep learning training system. In *Proceedings of the 11th USENIX symposium on operating systems design and implementation* (pp. 571–582). USENIX Association.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (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* (pp. 154–164). EMNLP-IJCNLP.

Guo, B., Zhang, C., Liu, J., & Ma, X. (2019). Improving text classification with weighted word embeddings via a multi-channel TextCNN model. *Neurocomputing*, *363*, 366–374.

Hazarika, D., Poria, S., Mihalcea, R., Cambria, E., & Zimmermann, R. (2018). ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2594–2604). ACL.

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L. P., & Zimmermann, R. (2018). 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* (pp. 2122–2132). NAACL.

Jiao, W., Lyu, M., & King, I. (2020). Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8002–8009). AAAI.

Jiao, W., Yang, H., King, I., & Lyu, M. R. (2019). HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 397–406). NAACL.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1746–1751). EMNLP.

Kim, M., & Kim, H. (2018). Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances. *Advances in Neural Information Processing Systems*, *101*, 1–5.

Kumar, H., Agarwal, A., Dasgupta, R., & Joshi, S. (2018). Dialogue act sequence labeling using hierarchical encoder with crf. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3440–3447). AAAI.

Li, X., Song, K., Feng, S., Wang, D., & Zhang, Y. (2018). A co-attention neural network model for emotion cause analysis with emotional context awareness. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4752–4757). EMNLP.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the eighth international joint conference on natural language processing* (pp. 986–995). IJCNLP.

Li, J., Xu, H., Deng, J., & Sun, X. (2016). Hyperbolic linear units for deep convolutional neural networks. In *2016 international joint conference on neural networks* (pp. 2161–4407). IEEE.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 6818–6825). AAAI.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Mohammad, S. (2018). Word effect intensities. In *Proceedings of the eleventh international conference on language resources and evaluation* (pp. 174–183). ELRA.

Molchanov, P., Tyree, S., Karras, T., Aila, T., & Kautz, J. (2017). Pruning convolutional neural networks for resource efficient inference. In *5th international conference on learning representations* (pp. 1–17). ICLR.

Olander, A., & Koinberg, I. (2017). Health dialogue as a tool for health promotion on individual, group and organisational levels. *British Journal of School Nursing*, *12*, 331–336.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). EMNLP.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Morency, Louis, P., & Zadeh, A. (2017). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 873–883). ACL.

Poria, S., Hazarika, D., Majumder, N., & Naik, G. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 527–536). IJCNLP.

Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *Social Sciences Information*, *7*, 100943–100953.

Qin, L., Che, W., Li, Y., Ni, M., & Liu, T. (2020). DCR-Net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8665–8672). AAAI.

Strapparava, C., & Valitutti, A. (2004). Wordnet affect: An affective extension of wordnet. In *2020 IEEE 8th international conference on smart city and informatization* (pp. 1083–1086). LREC.

Tan, G. W. H., Ooi, K. B., Leong, L. Y., & Lin, B. (2014). Predicting the drivers of behavioral intention to use mobile learning: A hybrid SEM-neural networks approach. *Computers in Human Behavior*, *36*, 198–213.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). MovieQA: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4631–4640). IEEE.

Upadhyay, P. K., & Nagpal, C. (2020). Wavelet based performance analysis of SVM and RBF Kernel for classifying stress conditions of sleep EEG. *Science and Technology*, *23*, 292–310.

Wu, L., Morstatter, F., & Liu, H. (2018). SlangSD: Building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, *52*, 839–852.

Ye, Y., Zhang, S., Li, Y., Qian, X., Tang, S., Pu, S., & Xiao, J. (2021). Video question answering via grounded cross-attention network learning. *Information Processing & Management*, *57*, 102265.

Young, T., Pandelea, V., Poria, S., & Cambria, E. (2020). Dialogue systems with audio context. *Neurocomputing*, *338*, 102–109.

Zhong, P., & Miao, C. (2019). Ntuer at SemEval-2019 Task 3: Emotion classification with word and sentence representations in RCNN. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 282–286). ACL.

Zhong, P., Wang, D., & Miao, C. (2019). Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 165–176). EMNLP-IJCNLP.

Zhou, L., Gao, J., Li, D., & Shum, H. Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, *46*, 53–93.