



MODE-Bi-GRU: orthogonal independent Bi-GRU model with multiscale feature extraction

Wei Wang^{1,2} · Wenhan Ruan^{2,3}  · Xiangfu Meng²

Received: 6 December 2022 / Accepted: 12 July 2023 / Published online: 9 October 2023
© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2023

Abstract

The core of sentence classification is to extract sentence semantic features. The existing hybrid methods have huge parameters and complex models. Due to the limited dataset, these methods are prone to feature redundancy and overfitting. To address this issue, this paper proposes an orthogonal independent Bi-GRU sentence classification model with multi-scale feature extraction, called Multi-scale Orthogonal Independent Bi-GRU (MODE-Bi-GRU). First, the hidden state of the Bi-GRU model is split into multiple small hidden states, and the corresponding recursive matrix is constrained orthogonally. Then, multiple sliding windows of different sizes are defined according to the forward and reverse angles of the sentence, and the sliding window is obtained. Finally, different sentence fragments are superimposed and input to the model, and the output results of multiple small Bi-GRU models are spliced and processed by soft pooling. The improved focal loss function is adopted to speed up the convergence of the model. Compared to the existing models, our proposed model achieves better results on four benchmark datasets, and it has better generalization ability with fewer parameters.

Keywords Feature redundancy · Generalization ability · Multi-scale feature extraction · Orthogonally constrain · Sentence classification

1 Introduction

Sentence classification is a foundational and conventional task in Natural Language Processing (NLP), and it is widely used in many subfields, such as text emotion analysis (Wang and Lai 2016; Perone 2018) and question classification (Shi and Tian 2016). The core problem of sentence classification is to understand text semantics by parsing key phrases located in different positions (Wang et al. 2015).

Responsible editor: Charalampos Tsourakakis.

Wenhan Ruan and Xiangfu Meng have contributed equally to this work.

Extended author information available on the last page of the article

Text Convolutional Neural Networks (TextCNN) are good at deriving n-gram characteristics of text through convolution operations, nonlinear layers, and pooling layers and have achieved splendid results in sentence classification (Kalchbrenner 2014; Kim 2014). However, the convolution operation is linear, which may not be adequate to remove non-sequential dependencies of phrases (Lei 2015) and may lose sequence information (Madasu 2019). In the sentence "The food is delicious though the restaurant is not almost as big.", the weighted sum of the second half "though the restaurant is not almost as small" does not capture the discontinuous dependency of "though not small" well and ignores the order information.

Long Short-Term Memory (LSTM) recurrent neural networks (Hochreiter 1997) and Gated Recurrent Unit (GRU) neural networks are appropriate for encoding structural correlations by reserving preceding word representations and sequential information semantics. However, when LSTM understands the words at time t , it can only obtain the information of 0 to $t-1$ but cannot obtain the information of words after time t . At the end of the complete sequence, in the above example, the sentence is still biased towards the following words, and the very first important word "delicious" is forgotten (Yin and Yu 2017). To address this issue, some current methods (Lai and Liu 2015; Wang 2016; Zhang 2016; Song 2018) combine CNN and LSTM by stacking or using bidirectional LSTM. However, simply stacking multiple layers easily results in feature redundancy and overfitting because only relatively small training sets are available for sentence classification tasks (Yin 2016; Guo 2019). Due to the complexity of the LSTM framework, the calculation is extremely complex, and the storage of redundant intermediate variables requires a lot of training time and memory space. LSTM and GRU can only use historical information to judge the current information but cannot use future information. Thus, they cannot make accurate judgments and fully extract sentence information.

Therefore, some studies (Zhao et al. 2018; Zhou 2018) use attention mechanisms to height prominent features and eliminate redundancy. However, attention mechanisms also increase the number of parameters, and training on limited datasets still suffers from overfitting.

In summary, it is necessary to explore a better sentence classification structure that balances generalization ability and complexity. This paper proposes a light-weight model Multiscale Orthogonal Independent Bidirectional GRU (MODE-Bi-GRU) fused with multi-scale feature extraction. The model reduces parameters, improves generalization ability, and considers different scaled n-gram semantic features. First, inspired by Zhang (2016), the hidden state of Bidirectional GRU (Bi-GRU) is unpacked into several independently updated small hidden states to reduce the number of parameters. Meanwhile, orthogonal constraints are placed on the recursive transition matrix of small hidden states, which improves a variety of features. This framework is called Bidirectional Orthogonal Independent GRU (ODE-Bi-GRU). Then, Bi-ODE-GRU is used within a local window to extract n-gram semantic features instead of simply using weighted sums like convolution. Especially, this paper introduces a three-size window operation that splits a sentence into multiple clauses in forward and reverse directions by sliding windows and stacks them together. These clauses are considered a mini-batch and can be processed in parallel by a shared ODE-Bi-GRU. In this paper, the last hidden state of

ODE-Bi-GRU is used as the n -gram semantic feature of each clause. Next, to capture variable-size phrases in the sentence, windows of various scales and different ODE-Bi-GRUs are used to extract features of multi-scale phrases. This framework is called bidirectional multi-scale ODE-GRU (MODE-Bi-GRU). Similar to CNN, MODE-Bi-GRU can derive multi-scale n -gram semantic features while reserving the nonlinearity and long-term dependencies of GRU, with strong modeling ability but fewer parameters. MODE-Bi-GRU is similar to a one-dimensional CNN using various filters with varied window sizes, but it uses recursive transformations instead of convolution operations. This paper conducts experiments on four sentence classification datasets, and the experimental results reveal that the model achieves comparable or better results than other models on datasets with fewer parameters.

2 Related work

The TextCNN-based model uses a deep CNN model with dynamic k -max pooling operations for semantic modeling of text. However, a simple single-layer CNN coupled with fine-tuned word embeddings can also achieve significant results. Some researchers also use various word embeddings as input to obtain further improvement.

A transformable CNN that can adaptively adjust the range of convolutional filters is proposed in Xiao and Chen (2018). Although methods based on the above excel are applied to local semantic feature extraction, the ability of the model to extract discontinuous dependencies and sequence information is limited by linear convolution operations.

The Recurrent Neural Network (RNN) model is appropriate for processing text sequences and modeling long-term dependencies, so it is also used for text modeling. Recently, some studies have added residual connections (Wang 2016) or dense connections (Ding and Yu 2018) to the recurrent structure to avoid vanishing gradients. A memory rotation unit (Dangovski 2019) was introduced to RNN to recall distant information.

HS-LSTM (Zhang 2018) can automatically build representations in a text through reinforcement learning. However, these RNN-based models still prefer to keep the words in the second half of the text and forget the words in the first half of the text. The model combining the bidirectional LSTM model and the attention mechanism can retain both the first half of the text and the second half of the text, but it cannot eliminate the problem of too long training time caused by complex calculations, and the training on limited data sets will suffer from overfitting.

To handle this problem, a common strategy is to combine the strengths of CNNs and RNNs through stacking. In Zhou and Liu (2015), max pooling is integrated into RNN to solve the bias problem of RNN. Phrase features are extracted using 1D convolutions, and then LSTMs are used to obtain text representations. Some subsequent work (Lee 2016) first uses LSTM to model long-term dependencies and then applies CNN to extract features of the text. However, these methods just stack multiple layers, resulting in feature redundancy and overfitting due to limited datasets. Some researchers introduced attention mechanisms (Er 2016; Lin et al. 2017) to strengthen

prominent features, but this causes a large number of parameters to overfit on small-scale datasets. A more flexible method is to replace the convolution operation with a tensor product or RNN unit. This method can directly capture nonlinear n-gram features, but it only considers n elements of a fixed scale.

The framework Dual-task Temporal Recurrent Reasoning Network for Joint Dialog Sentiment Classification (DARER) proposed at the ACL 2022 conference uses predictive-level interactions instead of semantic-level interactions to model explicit dependencies, so it is more consistent with human intuitive classification methods. (Xing 2022)

The most related works to our method are DRNN (Wang 2018) and MODE-LSTM (Ma and Yan 2020), which use RNN to learn semantic features. There are several differences: (1) DRNN uses GRU as a recursive unit; MODE-LSTM uses ODE-LSTM as a recursive unit and uses the Max pooling operation to obtain feature phrases representing sentence sentiment. This paper uses ODE-GRU and soft pooling operation to obtain overall features representing sentence sentiment, which achieves better generalization performance. (2) Instead of sequentially sliding to extract features, this paper introduces three windows of different sizes to extract features in parallel, which is faster than the DRNN algorithm. (3) Both MODE-LSTM and the model proposed in this paper consider multi-scale n-gram features of sentences, while DRNN only considers a fixed scale.

3 Model

3.1 Orthogonal independent Bi-GRU(ODE-Bi-GRU)

The vector related to an input sentence T is $\{X_1, X_2, \dots, X_T\}$, $X_i \in R^{d_0}$, where d_0 is the word embedding dimension of each word, and the LayerNormalization operation is performed. The hidden state $h_t \in R^d$ of each GRU unit is:

$$\begin{pmatrix} r_t \\ z_t \end{pmatrix} = \sigma \left(\begin{pmatrix} W_r \\ W_z \end{pmatrix} \cdot h_{t-1} + \begin{pmatrix} U_r \\ U_z \end{pmatrix} \cdot X_t \right) \quad (1)$$

$$O_t = \tanh(W_o \cdot (h_{t-1} * r_t) + U_o X_t) \quad (2)$$

$$h_t = O_t * z_t + (1 - z_t) * h_{t-1} \quad (3)$$

where r_t , z_t and O_t represent the update gate, reset gate and out gate of the GRU unit, respectively; $W_r, W_z, W_h, W_o \in R^{d \times d}$, $U_r, U_z, U_h, U_o \in R^{d_0}$ are the learnable parameters; X_t represents the sentence vector input at time t ; h_{t-1} is the hidden state at time $t-1$. The number of parameters of the GRU model is $3d \times (d + d_0)$, and the space complexity is $O(d^2)$. There is an overfitting problem for the dataset of limited sentence classification tasks.

To reduce the number of parameters, the GRU's hidden state h_t is equally divided into K independent hidden states, i.e., $\tilde{h}_t = [\tilde{h}_t^1, \dots, \tilde{h}_t^K]^T$, $\tilde{h}_t^i \in R^p$, $p = d/K$, and the

corresponding recursive matrix is $\tilde{W} = [\tilde{W}^1, \dots, \tilde{W}^K]$, $\tilde{W}^K \in R^{3p \times p}$. At time t of the sequence, each small hidden state \tilde{h}_t^i is individually updated by a separate recursive matrix \tilde{W}^i , and these \tilde{h}_t^i are connected and merged into \tilde{h}_t . This process can be expressed as the following formulas:

$$\begin{pmatrix} \tilde{r}_t \\ \tilde{z}_t \end{pmatrix} = \sigma \left(\begin{pmatrix} \tilde{W}_r \\ \tilde{W}_z \end{pmatrix} \otimes \tilde{h}_{t-1} + \begin{pmatrix} U_r \\ U_z \end{pmatrix} \cdot X_t \right) \tag{4}$$

$$\tilde{O}_t = \tanh(\tilde{W}_o \cdot (\tilde{h}_{t-1} * \tilde{r}_t) + U_o X_t) \tag{5}$$

$$\tilde{h}_t = \tilde{O}_t * \tilde{z}_t + (1 - \tilde{z}_t) * \tilde{h}_{t-1} \tag{6}$$

where \otimes is a vector point operation. For example,

$$\tilde{W} \otimes \tilde{h}_{t-1} = [\tilde{W}^1 \tilde{h}_{t-1}^1, \dots, \tilde{W}^K \tilde{h}_{t-1}^K]^\top \tag{7}$$

The ODE-Bi-GRU model is improved based on ODE-GRU, by adding a reverse input of sentence. Each hidden state in the reverse ODE-GRU model is expressed as:

$$\begin{pmatrix} \hat{r}_t \\ \hat{z}_t \end{pmatrix} = \sigma \left(\begin{pmatrix} \hat{W}_r \\ \hat{W}_z \end{pmatrix} \otimes \hat{h}_{t-1} + \begin{pmatrix} \hat{U}_r \\ \hat{U}_z \end{pmatrix} \cdot X_{T-t} \right) \tag{8}$$

$$\hat{O}_t = \tanh(\hat{W}_o \cdot (\hat{h}_{t-1} * \hat{r}_t) + \hat{U}_o X_{T-t}) \tag{9}$$

$$\hat{h}_t = \hat{O}_t * \hat{z}_t + (1 - \hat{z}_t) * \hat{h}_{t-1} \tag{10}$$

The hidden states in ODE-Bi-GRU are:

$$h_t = w_t * \tilde{h}_t + v_t * \hat{h}_t \tag{11}$$

where w_t, v_t are learnable parameters, and this paper only keeps the last hidden state (return_sequence=False).

In the derivation of the recurrent neural network, there are multiple parameter matrices (collectively referred to as W), as shown in Equation 12:

$$\frac{\partial f_n}{\partial W} = \frac{\partial f_n}{\partial h_n} f_{n-1} + W \frac{\partial f_n}{\partial h_n} \frac{\partial f_{n-1}}{\partial h_{n-1}} f_{n-2} + \dots + W^{(n-1)} \frac{\partial f_n}{\partial h_n} \frac{\partial f_{n-1}}{\partial h_{n-1}} \dots \frac{\partial f_1}{\partial h_1} \frac{\partial h_1}{\partial W} \tag{12}$$

If the eigenvalue λ of the parameter matrix W is less than 1, the gradient disappears after continuous multiplication; if $|\lambda|$ is greater than 1, the continuous multiplication causes the gradient to explode. If the parameter matrix is initialized as an orthogonal matrix, and the eigenvalue's modulo of the orthogonal matrix is 1, gradient explosion or gradient disappearance can be avoided. So, it is necessary to initialize the parameter matrix of these small GRUs as an orthogonal matrix.

3.2 Equipping ODE-Bi-GRU with sliding window

The core of the sentence classification task is to analyze the semantics of keywords and variable-size phrases in the sentence. Although CNNs can extract n-gram features, linear convolution operations are inadequate to model the sequential information and discontinuous dependencies of a sentence. One-way LSTM focuses on retaining the content of the latter part of the sentence, and it is easy to forget the content of the former part of the sentence. The ODE-Bi-GRU in this paper can preserve word order and control the preservation or forgetting of information through bidirectional gates to model discontinuous dependencies. Taking the sentence "The food is delicious though the restaurant is not almost as big" as an example, ODE-LSTM still focuses on retaining the weight information of "not" and "big", through the one-way gate control unit. By contrast, ODE-Bi-GRU uses the forward gate control unit to selectively and gradually retain the semantic information of "not" and "big", while gradually forgetting other words; meanwhile, it uses the reverse gate control unit to retain the semantic information of "delicious" and "though", while gradually forgetting other words.

Therefore, this paper uses a sliding window for ODE-Bi-GRU to extract n-gram features. In this case, the cyclic transformation of ODE-Bi-GRU can only be accomplished in a local window and scales with the sentence. For the sliding window position corresponding to time series t , ODE-Bi-GRU will sequentially process consecutive words within the forward sentence and reverse sentence range of ($emph t-S+1, t$) and generate related hidden states. The output of ODE-Bi-GRU is used as the n-gram feature in the scanning range of the sliding window:

$$h_t = \text{ODE-Bi-GRU}([X_{t-S+1}, \dots, X_t], [X_t, \dots, X_{t-S+1}]) \quad (13)$$

Then, the dimension of the vector is restored to d . Meanwhile, $s-1$ 0s are added to the front of the sentence to ensure that the window size of each time series is the same (as shown in Fig. 1b). Such a local method is similar to DRNN, but it processes all windows sequentially, which is very time-consuming. However, it is found that all windows are independent of each other, so these windows can be processed in parallel with GPU to enhance computational efficiency.

This paper also introduces the triple-S (Sliding Split Stacking) operation to combine all windows, as illustrated in Fig. 1b. First, a forward sentence and the corresponding reverse sentence are divided into multiple clauses through a sliding window of length S , and then they are stacked together, where $B_1 \in R^{T \times S \times d_0}$, $B_2 \in R^{T \times S \times d_0}$.

3.3 Multi-scale ODE-Bi-GRU (MODE-Bi-GRU)

Sentence phrases have n-gram features of different scales, which makes the use of only a fixed window not sufficient. Therefore, it is necessary to use windows of different scales to extract n-gram features of different scales in parallel with ODE-Bi-GRU. The structure of the bidirectional multi-scale ODE-GRU (MODE-Bi-GRU) model is shown in Fig. 1a. Based on the sliding stacking operation of windows

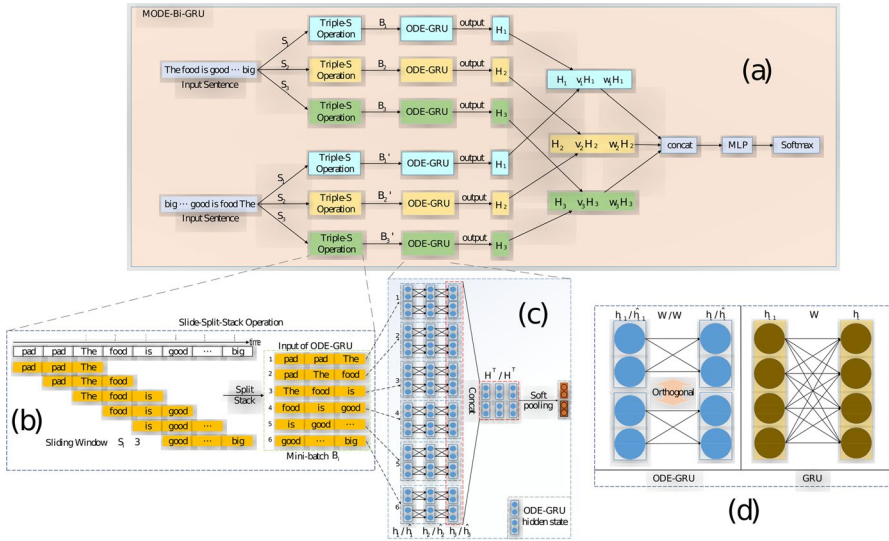


Fig. 1 **a** The structure of MODE-Bi-GRU with three windows of different scales $[S_1, S_2, S_3]$. The input sentence is transformed into three mini-batches $[B_1, B_2, B_3]$ by the Triple-S operation. These mini-batches are respectively fed into different initialized ODE-GRUs to derive n-gram features for each scale. **b** The detail of Triple-S operation. **c** The procedure of performing mini-batch for an ODE-GRU. **d** The comparison of ODE-GRU and GRU. Here, ODE-GRU disentangles the hidden state into two small hidden states. An orthogonal limitation is applied to the recurrent matrix \tilde{W}/\hat{W} to improve the diversity of features (Color figure online)

$[S_1, \dots, S_m]$ of different scales (S_i is the window size), the sentence is transformed into multiple mini-batches $B = [B_1, \dots, B_m]$ and $B' = [B'_1, \dots, B'_m]$. Then B and B' are input into different ODE-GRUs to obtain the n-gram feature matrix:

$$H_{S_m} = [h_{S_m,1}, \dots, h_{S_m,t}, \dots, h_{S_m,T}]^T \tag{14}$$

$h_{S_m,t} \in R^d$ represents time series t , the forward and reverse n-gram features scanned by the sliding window.

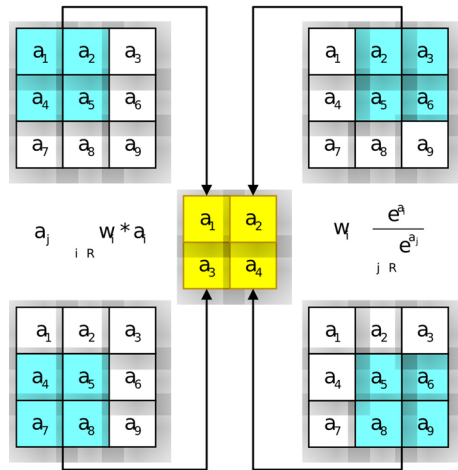
$$h_t = \text{ODE-Bi-GRU}_{S_m} ([X_{t-S_m+1}, \dots, X_t], [X_t, \dots, X_{t-S_m+1}]) \tag{15}$$

In each n-gram feature matrix along the T-dimension direction, each vector in the T-dimension direction is a vector represents of a sentence. The significant features corresponding to each scale, are extracted through soft pooling (SP), as shown in Fig. 2. These features are spliced to form a multi-scale feature expression:

$$F = [SP(H_{S_1}), \dots, SP(H_{S_m})]^T \tag{16}$$

The feature representation F is reconstructed into a vector and sent to the MLP layer. Then, it is processed by the activation function ReLU and the softmax layer for classification.

Fig. 2 Soft pooling, where a_i is the feature extracted for each ODE-Bi-GRU (Color figure online)



3.4 Loss function

The loss function corresponding to the MODE-Bi-GRU model is an improved Focal loss (Lin and Girshick 2017) function that adds a penalty term to the Focal loss function.

The sentence classification model is a probability model, and the corresponding loss function is the cross entropy loss function:

$$Loss = \frac{1}{N} \sum_{i=1}^N -[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \tag{17}$$

where, N is the number of samples, y_i is the label corresponding to the sample, and p_i is the probability value predicted by the model. In the sentence classification dataset, there is a problem of unbalanced sample categories. The critical information provided by the small number of samples cannot be used in training to derive a loss value that can provide a correct guidance for model training. So this paper uses the Focal_loss function to solve this problem, and the specific mathematical formula is as follows:

$$Focal_Loss = \alpha \times (1 - p_i)^{\wedge G} \times Loss \tag{18}$$

where α and G are hyperparameters. $G = 2$ and $\alpha = 0.25$ lead to the best performance (Ma and Yan 2020). To ensure that each small Bi-GRU model learns different knowledge, a penalty term is introduced is as follows:

$$W = \left(\tilde{W} + \hat{W} \right) / 2 \tag{19}$$

$$LP = \sum_{i=1}^K \sum_{j=1}^K \|WW^T - I\|_2^2 \tag{20}$$

where I is the identity matrix. Assume that any two rows in W (the results of any two small Bi-GRUs) consider completely different aspects. According to the properties of orthogonal matrices, the dot product is 0 and $Lp = 1$. During training, the value of L is continuously reduced to ensure that different small Bi-GRUs learn different contents.

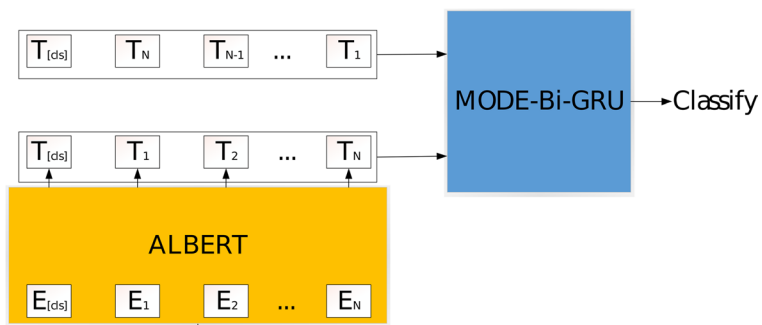
Therefore, the final loss function of the model is as follows:

$$L = Focal_Loss + \lambda \sum_{i=1}^m Lp_i \tag{21}$$

3.5 Combine with ALBERT

Recently, the pre-trained language model BERT (Devlin and Lee 2018) has been proven to be more effective than traditional word embeddings in fine-tuning downstream tasks. ALBERT (Lan et al. 2019) is improved on BERT, by factoring word embeddings and sharing parameters across layers. Though its parameter size is reduced from 110 to 12 M, the model representation capability is almost the same as that of BERT. Compared to word embeddings, ALBERT can learn context-sensitive sentence representations. However, recent work has shown that ALBERT’s self-attention mechanism distracts the text, thus ignoring important adjacent elements and phrases.

MODE-Bi-GRU can extract multi-scale local features bidirectionally, which is complementary to ALBERT representation. Therefore, this paper combines MODE-Bi-GRU with ALBERT to further improve the generalization performance of the model (Fig. 3). Specifically, the sentences are fed into the ALBERT, and the hidden representation of the ALBERT’s last layer is used as the input of MODE-Bi-GRU instead



The food tastes delicious though the restaurant is not almost as big

Fig. 3 The diagram of MODE-Bi-GRU fusing with ALBERT (Color figure online)

of GloVe and character embeddings. ALBERT provides contextualized sentence-level representation, which helps MODE-Bi-GRU to understand sentence semantics more accurately.

4 Experiments

4.1 Experimental setup

To evaluate the effectiveness of our proposed model, sentence classification experiments are conducted on four extensively studied datasets. The statistics of these datasets are listed in Table 1, where C is the number of target categories; L is the average length of the sentence; ML is the maximum length of the sentence, and CV is ten-fold cross-validation. These datasets come from different topics such as sentiment analysis, movie reviews (MR, SST2, SST5), and question type (TREC) classification. The Amazon dataset contains more than 1.3 million reviews, and each review contains information such as rating, title, review text, review time, and product ID. These reviews come from goods of different categories on the Amazon website, such as books, movies, music, electronics, etc. These reviews can be used for many natural language processing tasks, such as sentiment analysis, text classification, entity recognition, etc. The Yelp dataset contains reviews and ratings of businesses on the Yelp business social media site. The dataset contains more than 6 million reviews and 2 million pieces of business information, and each review contains information such as rating, comment text, comment time, user ID, and business ID. These reviews come from businesses of different types in different cities, such as restaurants, bars, beauty salons, etc. The ratios of positive and negative samples in the SST2 and Yelp.P datasets are 1.09:1 and 1.17:1, respectively.

In this paper, accuracy, recall, and F1-score are used to measure the performance of the sentence classification model. The calculations formulas are shown below::

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (22)$$

$$Recall = \frac{TP}{TP + FN} \quad (23)$$

$$Precision = \frac{TP}{TP + FP} \quad (24)$$

Table 1 Statistics of four dataset

Dataset	C	L	ML	Train	Dev	Test
MR	2	19	53	10662	–	CV
TREC	6	10	33	5452	–	500
SST2	2	19	53	6920	872	1821
SST5	5	18	53	8544	1101	2210

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (25)$$

where TP means the number of positive samples that are also predicted as positive; TN the number of negative samples that are also predicted as negative; FP means the number of negative samples that are predicted as positive; FN means the number of positive samples that are predicted as negative.

Since F1 score and recall are suitable for the research of binary classification problems, F1 score and recall are only added to the binary classification data sets SST2, MR and Yelp.P.

4.2 Experimental details

The word embedding is initialized by the glove pre-trained word vector published by Stanford. The vector has a dimension of 300, and it is merged with a 50d character embedding constructed by a convolutional layer with a max-pooling layer to avoid the problem that words cannot be found in the glove. In the experiment, the continuous debugging of the two parameters of the sliding window size and the number of GRU splits K reveals that when the sliding window size is [6,11,16] and K is 2, the effect is the best. The parameters corresponding to the model are shown in Table 2.

4.3 Baseline models

MODE-Bi-GRU is compared with three types of strong baselines:

- (1) CNN/RNN-based model: TextCNN, LSTM, GRU, Bi-LSTM and Bi-GRU.
- (2) Hybrid model: C-RNN directly fuses LSTM with RNN, DRNN, and self-attention mechanism (Vaswani et al. 2017).

Table 2 Model parameter setting

Parameter	Value
K	2
Hidden layer state dimension	50
Regularization	L2
Embedding dropout	0.2
MLP dropout	0.5
Sliding window	[6,11,16]
Optimizer	Adam
Learning rate	0.001
Batch size	60
α (Formula 17)	0.25
GAMA(G in Formula 17)	2
λ (Formula 20)	0.01

- (3) Capsule network: The HAC (Zheng and Wan 2019) model is complex, and each layer of the network is composed of deep dilated convolution and a capsule module.
- (4) MODE-LSTM: It was proposed at the EMNLP2020 conference, and it uses max pooling to perform pooling operations on the extracted features. In addition to the above models, this paper also uses ODE-Bi-GRU as a benchmark. The value of K is set to 6; so that the number of parameters is consistent with MODE-Bi-GRU.
- (5) MODE-GRU: It is compared with the unidirectional model to further verify the effectiveness of multi-scale independent orthogonality (control variable).
- (6) DARER: It was proposed by ACL in 2022, and the main architecture of this model is also based on Bi-LSTM.
- (7) MODE-Bi-GRU fused with ALBERT (MODE-Bi-GRU+ALBERT_{base}) is compared with some recent baseline models that are also combined with pre-trained sentence representations, including InferSent (Conneau and Schwenk 2017), HAC+ELMo, and MODE-LSTM+ALBERT_{base}.

4.4 Experimental results and analysis

The performance comparison results are presented in Table 3. (1) With fewer parameters, MODE-Bi-GRU achieves significantly better performance than DRNN, and the average accuracy is increased by more than 1.0%, which is 1.2% higher than that of MODE-LSTM. Compared with GRU, MODE-Bi-GRU can retain the features before and after the sentence and consider the features for the sentence of different scales, so MODE-Bi-GRU performs better than DRNN and MODE-LSTM.

(2) MODE-Bi-GRU is similar to the single-layer TextCNN model. It is simple and effective and achieves an average accuracy improvement of 1.9% over the recently proposed HAC model. Due to the reduction of model parameters, MODE-Bi-GRU has better generalization ability and achieves better performance than HAC in the test set.

(3) Although the parameters of TextCNN are less than MODE-Bi-GRU, its parameter size will increase with the filter window size. By contrast, the parameter size of MODE-Bi-GRU is independent of the window size.

(4) The average accuracy of Bi-ODE-GRU is 2.5%, 3.6%, 4.1% higher than that of Bi-GRU, which indicates the effectiveness of orthogonal and independent hidden states. The average accuracy, recall, and F1-score of MODE-Bi-GRU is 0.78%, 0.94%, 2.4% higher than that of ODE-Bi-GRU. It can be seen that the effect of fusing windows of different scales is better than that of fusing single-scale windows, which can prove the effectiveness of multi-scale feature extraction.

(5) By replacing the Glove word vector representation with the ALBERT representation, MODE-Bi-GRU can further improve the generalization performance, as shown in the last row in Table 4. Although the strong performance of ALBERT has been verified on a large number of datasets, it may tend to ignore local phrase information tends to be ignored due to the self-attention mechanism. Therefore, the combination of MODE-Bi-GRU and ALBERT can further improve the prediction

Table 3 Comparative experiment on four datasets

Model	#Params	MR Acc/Re/F ₁	TREC Acc/Re/F ₁	SST2 Acc/Re/F ₁	SST5 Acc/Re/F ₁	Amz.F Acc/Re/F ₁	Yelp.P Acc/Re/F ₁	Average Training Time(hour)
HAC	–	81.4/-/-	93.5/-/-	81.0/-/-	47.3/-/-	–	–	6.8
TextCNN	466K	80.2/-/-	92.0/-/-	80.6/-/-	46.8/-/-	–	–	3.6
LSTM	827K	80.3/-/-	93.1/-/-	80.7/-/-	47.8/-/-	–	–	3.9
GRU	620K	81.6/-/-	92.6/-/-	80.9/-/-	46.1/-/-	–	–	3.7
CNN-LSTM	1.1M	78.9/-/-	93.0/-/-	80.8/-/-	47.3/-/-	–	–	5.3
Self-Attention	42 M	80.3/-/-	92.9/-/-	78.6/-/-	46.2/-/-	–	–	10
MODE-LSTM	527K	81.4/76.6/83.3	94.0/-/-	81.3/76.5/77.1	48.1/-/-	58.6/-/-	89.3/88.7/89.6	3.6
MODE-GRU	385K	82.0/79.3/84.6	94.3/-/-	82.2/79.6/83.5	48.4/-/-	58.6/-/-	89.7/88.6/90.6	3.4
Bi-LSTM	1.6M	80.3/-/-	92.4/-/-	79.8/-/-	46.3/-/-	54.2/-/-	82.1/-/-	5.4
Bi-GRU	1.2M	80.7/-/-	92.8/-/-	78.3/-/-	46.4/-/-	54.5/-/-	83.2/-/-	4.9
DARER	–	80.6/79.3/81.1	93.8/-/-	81.2/80.7/81.6	47.1/-/-	56.4/-/-	87.6/87.2/88.3	4.8
ODE-Bi-GRU	–	81.4/-/-	93.5/-/-	81.0/-/-	47.3/-/-	–	–	4.2
MODE-Bi-GRU	790K	82.5/81.2/86.4	94.6/-/-	82.3/81.9/83.1	48.9/-/-	59.9/-/-	91.5/90.7/92.3	4.2
Combined with pre-trained language models								
InferSent	–	81.1/-/-	88.2/-/-	84.6/-/-	47.2/-/-	59.8/-/-	85.4/85.1/85.7	6.1
HAC+ELMo	–	84.7/-/-	96.4/-/-	89.2/-/-	48.5/-/-	61.2/-/-	86.7/86.2/87.3	5.7
ALBERT _{base}	12 M	85.8/85.2/86.1	94.7/-/-	90.3/87.5/91.8	50.7/-/-	60.2/-/-	92.4/91.8/92.7	5.8
BERT _{base}	110 M	86.3/85.2/87.4	95.3/-/-	92.5/92.2/92.8	52.7/-/-	61.6/-/-	93.8/93.6/94.2	20
MODE-LSTM+ALBERT _{base}	12.5M	86.9/86.4/87.1	96.1/-/-	93.2/92.9/93.3	52.9/-/-	66.4/-/-	94.1/92.7/94.6	6
MODE-Bi-GRU+ALBERT _{base}	12.7M	88.1/87.3/89.4	97.6/-/-	94.1/93.8/96.7	54.2/-/-	78.2/-/-	97.8/96.5/97.9	6.2

Bold values indicate the best performance

Table 4 Ablation experiment

Windows	Penalization	SST2	MR	TREC
[6, 11, 16]	✓	82.3	82.5	94.6
[6, 6, 16]	✓	82.0	82.1	93.5
[6, 16, 16]	✓	81.9	81.6	93.7
[6, 11, 16, 24]	✓	81.6	81.7	94.2
[6, 11, 16]	×	82.1	82.1	94.3

ability, which indicates that the MODE-Bi-GRU model can better understand semantics. It is worth noting that the classification performance of the MODE-Bi-GRU model has surpassed that of the pre-trained model InferSent. Also, the classification performance of the MODE-Bi-GRU+ALBERT_{base} model exceeds that of 110 M BERT_{base}, and the effectiveness and generalization of the MODE-Bi-GRU model are verified.

(6) According to the ablation experiment in Table 4, the effect of fusing the windows of three different scales is the best, and adding a penalty term to the objective function can improve the performance of the model.

To investigate how MODE-Bi-GRU differs from other models, its convergence trend is visualized in Fig. 4. It can be observed that the directly stacked C-LSTM (dark blue line) converges quickly on the training set but performs poorly on the validation or test set. Although the self-attention mechanism model (dark green line) can reduce feature redundancy, overfitting still occurs due to a large number of parameters. MODE-Bi-GRU (red line) has more superior generalization ability than other models on the dev or test set.

4.5 Case study

To explore why Bi-MODE-GRU outperforms TextCNN, DRNN, and MODE-LSTM, several of the most contributing positions in the largest pool are presented through visualization techniques. Four examples are selected from the SST dataset (these examples are biased towards aspect-level sentiment) and converted into different expressions. In the four cases shown in Table 5, TextCNN misses the key phrase "not". Therefore, the sentence is incorrectly classified as a negative sentence (although examples 2 and 3 are negative sentiments, but not judged by the word "not").

In the third example, MODE-LSTM and MODE-Bi-GRU capture discontinuous dependencies based on the keywords "but" and "awful". Therefore, they pay attention to the second half of the sentence for accurate classification. In the first example, both MODE-LSTM and MODE-Bi-GRU extract key phrases with "not splendidly decorated" discontinuous dependencies, but MODE-LSTM is unidirectional, and max pooling is applied to the extracted features. The transformation operation retains only one feature and still forgets the initial keyword "delicious". Therefore, MODE-LSTM fails in this case. However, MODE-Bi-GRU can extract multiple features through bidirectional multi-scale features. Meanwhile, it uses soft pooling to

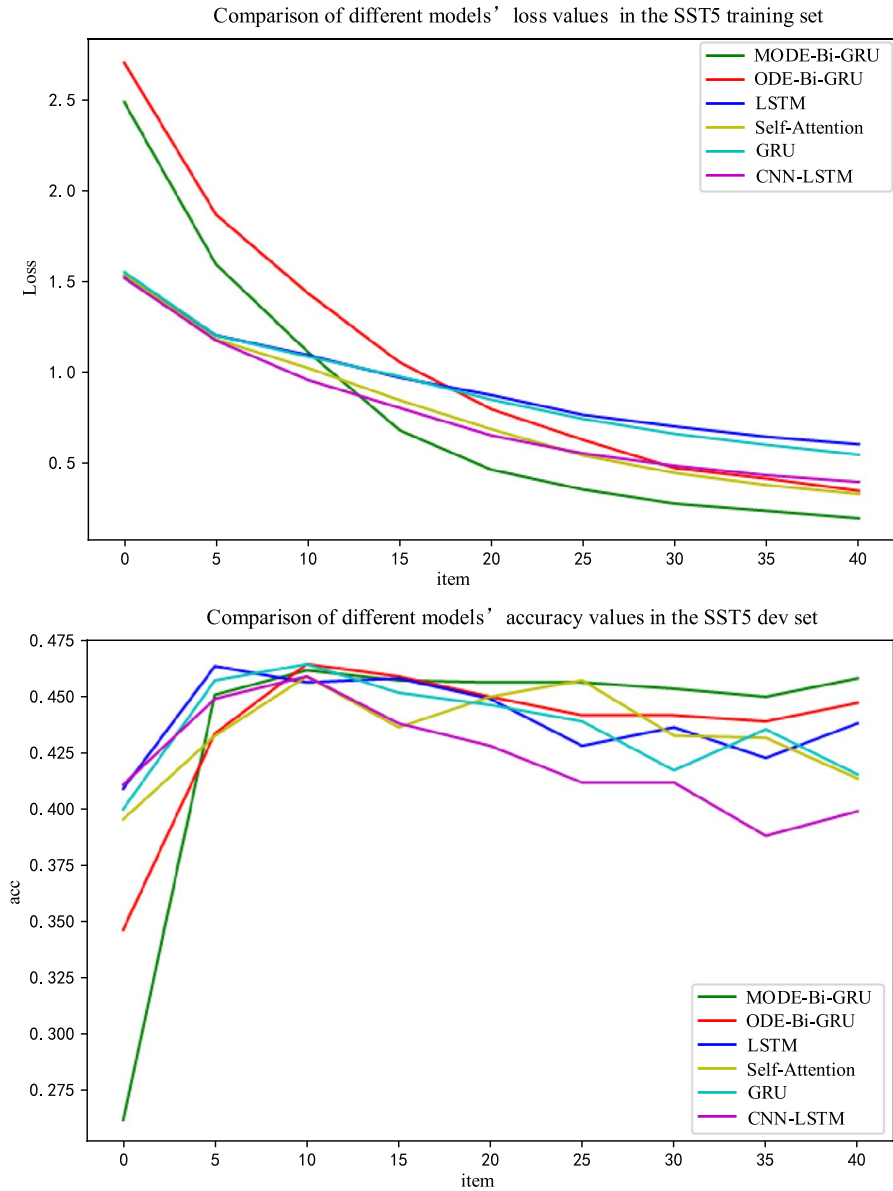


Fig. 4 The convergence analysis on SST2 datasets (Color figure online)

Table 5 Case study of our model compared to other models

Num	Examples	Lable	TextCNN	MODE-LSTM	MODE-Bi-GRU
1	The food tastes <u>delicious</u> though the restaurant is not almost as <u>splendidly decorated</u> . This restaurant is located <u>next</u> to the <u>waterfall</u> .	P	N	P	P
2	The food tastes <u>awful</u> though the restaurant is not almost as <u>poorly decorated</u> . This restaurant is located <u>next</u> to the <u>waterfall</u> .	N	N	P	N
3	This restaurant is located next to the waterfall. The restaurant is not almost as <u>poorly decorated</u> <u>but</u> the food tastes <u>awful</u> .	N	N	N	N
4	The restaurant is not almost as <u>splendidly decorated</u> <u>but</u> the food tastes <u>delicious</u> . This restaurant is located next to the waterfall.	P	N	P	P

Bold fonts, italics and underlines correspond to the important positions extracted by TextCNN, MODE-LSTM, and MODE-Bi-GRU respectively

accumulate multiple feature weights to obtain the main features of the sentence, thus obtaining the correct answer.

5 Conclusion

This study proposes a new parameter-efficient model, multi-scale n-gram semantic features from sentences. Different from the complex operations of stacking CNNs and RNNs or attaching too many parameters to the attention mechanism, our work provides a lightweight method to improve sentence classification. By decomposing the hidden state into multiple small orthogonal independent hidden states and using multiple sliding windows of different scales, MODE-Bi-GRU performs better than the popular CNN/RNN-based approach on different benchmark datasets. In future work, the effectiveness of the proposed model in aspect-level sentiment classification will be confirmed through large-scale validations.

Small sample learning (few-shot) technology is becoming mature. Smaller datasets such as SST and MR still contain more than a few thousand sentences, and small sample learning can use pre-trained models (even be trained on the CPU) to obtain better results on dozens of sentences, which greatly reduces the training cost and training time. The next step is to fuse the MODE-Bi-GRU+ALBERT_{base} model with the prompt method.

Declarations

Conflict of interest Authors have no conflict of interest to declare.

References

- Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2017) Supervised learning of universal sentence representations from natural language inference data. arXiv preprint [arXiv:1705.02364](https://arxiv.org/abs/1705.02364)
- Dangovski R, Jing L, Nakov P, Tatalović M, Soljačić M (2019) Rotational unit of memory: a novel representation unit for rnns with scalable applications. *Trans Assoc Comput Ling* 7:121–138
- Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
- Ding Z, Xia R, Yu J, Li X, Yang J (2018) Densely connected bidirectional lstm with applications to sentence classification. In: *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 278–287. Springer
- Er MJ, Zhang Y, Wang N, Pratama M (2016) Attention pooling-based convolutional neural network for sentence modelling. *Inf Sci* 373:388–403
- Guo H, Mao Y, Zhang R (2019) Augmenting data with mixup for sentence classification: an empirical study. arXiv preprint [arXiv:1905.08941](https://arxiv.org/abs/1905.08941)
- Hochreiter Schmidhuber (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
- Kalchbrenner N, Grefenstette E, Blunsom PA (2014) A convolutional neural network for modelling sentences. arXiv preprint [arXiv:1404.2188](https://arxiv.org/abs/1404.2188)
- Kim Y (2014) Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1746–1751


- Lai S, Xu L, Liu K, Zhao J (2015) Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI Conference on Artificial Intelligence
- Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: A lite bert for self-supervised learning of language representations. arXiv preprint [arXiv:1909.11942](https://arxiv.org/abs/1909.11942)
- Lee JY, Dernoncourt F (2016) Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint [arXiv:1603.03827](https://arxiv.org/abs/1603.03827)
- Lei T, Barzilay R, Jaakkola T (2015) Molding cnns for text: non-linear, non-consecutive convolutions. arXiv preprint [arXiv:1508.04112](https://arxiv.org/abs/1508.04112)
- Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988
- Lin Z, Feng M, Santos CND, Yu M, Xiang B, Zhou B, Bengio Y (2017) A structured self-attentive sentence embedding. arXiv preprint [arXiv:1703.03130](https://arxiv.org/abs/1703.03130)
- Madasu A, Rao VA (2019) Sequential learning of convolutional features for effective text classification. arXiv preprint [arXiv:1909.00080](https://arxiv.org/abs/1909.00080)
- Ma Q, Lin Z, Yan J, Chen Z, Yu L (2020) Mode-lstm: a parameter-efficient recurrent network with multi-scale for sentence classification. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6705–6715
- Perone C, Silveira R, Paula TS (2018) Evaluation of sentence embeddings in downstream and linguistic probing tasks. arXiv preprint [arXiv:1806.06259](https://arxiv.org/abs/1806.06259)
- Shi Y, Yao K, Tian L, Jiang D (2016) Deep lstm based feature mapping for query classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1501–1511
- Song X, Petrak J, Roberts A (2018) A deep neural network sentence level classification method with context information. arXiv preprint [arXiv:1809.00934](https://arxiv.org/abs/1809.00934)
- Vaswani Shazeer, Parmar, Uszkoreit Jones, Gomez Kaiser, Polosukhin: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
- Wang B (2018) Disconnected recurrent neural networks for text categorization. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 2311–2320
- Wang X, Jiang W, Luo Z (2016) Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2428–2437
- Wang Y, Tian F (2016) Recurrent residual learning for sequence classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 938–943
- Wang P, Xu J, Xu B, Liu C, Zhang H, Wang F, Hao H (2015) Semantic clustering and convolutional neural network for short text categorization. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 2, pp. 352–357
- Wang J, Yu LC, Lai KR, Zhang X Dimensional sentiment analysis using a regional CNN-LSTM model. In: Proceedings of the 54th Annual Meeting of the Association for Computational, vol. 2, pp. 225–230. Association for Computational Linguistics, Berlin, Germany (2016)
- Xiao L, Zhang H, Chen W, Wang Y, Jin J (2018) Transformable convolutional neural network for text classification. In: IJCAI, pp. 4496–4502
- Xing B, Ivor IW (2022) Darer: Dual-task temporal relational recurrent reasoning network for joint dialog sentiment classification and act recognition. arXiv preprint [arXiv:2203.03856](https://arxiv.org/abs/2203.03856)
- Yin W, Kann K, Yu M, Schütze H (2017) Comparative study of cnn and rnn for natural language processing. arXiv preprint [arXiv:1702.01923](https://arxiv.org/abs/1702.01923)
- Yin W, Schütze H (2016) Multichannel variable-size convolution for sentence classification. arXiv preprint [arXiv:1603.04513](https://arxiv.org/abs/1603.04513)
- Zhang T, Huang M, Zhao L (2018) Learning structured representation for text classification via reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32
- Zhang R, Lee H, Radev D (2016) Dependency sensitive convolutional neural networks for modeling sentences and documents. arXiv preprint [arXiv:1611.02361](https://arxiv.org/abs/1611.02361)
- Zhang Y, Roller S, Wallace B (2016) Mgnc-cnn: A simple approach to exploiting multiple word embeddings for sentence classification. arXiv preprint [arXiv:1603.00968](https://arxiv.org/abs/1603.00968)
- Zhao J, Zhan Z, Yang Q, Zhang Y, Hu C, Li Z, Zhang L, He Z (2018) Adaptive learning of local semantic and global structure representations for text classification. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2033–2043

- Zheng W, Zheng Z, Wan H, Chen C (2019) Dynamically route hierarchical structure representation to attentive capsule for text classification. In: IJCAI, pp. 5464–5470
- Zhou C, Sun C, Liu Z, Lau F (2015) A c-lstm neural network for text classification. arXiv preprint [arXiv:1511.08630](https://arxiv.org/abs/1511.08630)
- Zhou Q, Wang X, Dong X (2018) Differentiated attentive representation learning for sentence classification. In: IJCAI, pp. 4630–4636

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Wei Wang^{1,2} · Wenhan Ruan^{2,3}  · Xiangfu Meng²

✉ Wenhan Ruan
ruanzong0427@gmail.com

Wei Wang
ww431894@163.com

Xiangfu Meng
marxi@126.com

¹ Department of Foundation, Liaoning Technical University, Longwan South Street, Longwan South Huludao, Liaoning 125105, China

² School of Electronic and Information Engineering, Liaoning Technical University, Street, Huludao, Liaoning 125105, China

³ Shenyang Hunnan Branch, China Construction Bank, Xinlong Street, Shenyang 110179, Liaoning, China