J Inf Process Syst, Vol.20, No.6, pp.758~766, December 2024
https://doi.org/10.3745/JIPS.02.0221
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Chinese Long-Text Classification Strategy Based on Fusion Features

Li Lin, Yidan Wang, Yan Wang, and Changhua Tang*

Abstract

In the process of Chinese long-text classification, due to the large amount of text data and complex features, methods suitable for ordinary text classification often lack sufficient accuracy, which directly leads to frequent classification failures in long-text environments. To solve this problem, the research designed a bi-directional long short-term memory (Bi-LSTM) model that combines forward and backward operations and utilized attention mechanisms to improve fusion. At the same time, the bi-directional encoder representations from transformers (BERT) model was introduced into the text processing to form a long-text classification model. Finally, different datasets were tested to verify the actual classification effect of the model. The research results showed that under different dataset environments, the classification accuracy rates of the designed models were 92.93% and 93.77%, respectively, which are the models with the highest classification accuracy rates among the same type of models. The calculation time was 85.42 seconds and 117.51 seconds, respectively, which are the models with the shortest calculation time among the same type of models. It can be seen that the research designed long-text classification model innovatively combined the BERT model, convolutional neural network model, Bi-LSTM model, and attention mechanism structure based on the data characteristics of long-text classification, enabling the model to achieve higher classification accuracy in a shorter computational time. Moreover, it has better classification results in actual long-text classification, overcomes the classification failure problem caused by complex text features in the long-text classification environment, and provides a possibility for long-text specific classification paths.

Keywords

Attention Mechanism, BERT Model, Fused Features, Neural Network

1. Introduction

In the process of modernization, the management of a large amount of Chinese text data on the network has become a major issue. More efficient text classification management for different types of text can improve the speed of online text retrieval, lay the foundation for further network data analysis, and provide assistance in technical fields such as text reading, automatic text recognition, computer chat sentence recognition, etc. [1-3]. Although traditional machine learning text classification methods can achieve certain text classification results, the limitations of their feature extraction performance lead to higher classification performance only in regular text and short text. When they face long-text data with more text features, the classification accuracy and efficiency of classification operations will be

Manuscript received February 23, 2023; first revision March 28, 2023; second revision May 30, 2023; third revision August 3, 2023; accepted August 6, 2023.

* **Corresponding Author:** Changhua Tang (tangch564@163.com)
School of Computer Science and Engineering, College of Humanities and Information Changchun University of Technology, Changchun, China (linli706@163.com, wydmn0928@163.com, ronaldo6688@sohu.com, tangch564@163.com)
Current affiliation for Li Lin, College of Engineering and Technology, Jilin Agricultural University, Jilin, China.
Current affiliation for Yidan Wang, School of Mechanical and Electrical Engineering, Changchun University of Technology, Changchun, China.

significantly reduced [4-6]. Therefore, there is a need for an efficient text classification method that can provide the most basic information classification processing capabilities for computer deep semantic recognition technology, and lay the foundation for computer text semantic analysis technology and deep data analysis technology.

2. Related Works

The research in the field of text classification is gradually intensifying and different approaches are being applied in this field. Yao et al. [7] developed a model for the text classification problem, which is based on a single-production corpus to build a text graph. It shows the text graph network model can effectively improve the speed of text division. Akhter et al. [8] designed a machine learning long text classification method for multi-format, multi-purpose text datasets. The results showed that the division accuracy of this program was 91.8%, 95.4%, and 93.3% on large, medium, and small datasets, respectively, which had superior performance advantage compared with similar models. Liu and Guo [9] employed the LSTM pattern to form a network that can extract higher-level feature representation from word vectors. The model was able to extract higher-level feature representations from word vectors and apply them to contextual representations. Finally, a softmax classifier was applied to process the contextual data. The results showed that the pattern captured local features of phrases more accurately and had higher accuracy in overall classification. Huan et al. [10] combined multi-scale neural networks with bi-directional networks to form a new text classification model. Hybrid attention was applied to extract deep semantics and deep semantics were combined with shallow semantics to form a more comprehensive and accurate classification model. The results showed that the program could significantly improve the classification of text. The authors of [11] designed a hybrid recurrent attention network, which combined a long-short term neural network with a network to catch contextual data information in both directions and judged the long-term dependency of long text to capture words with higher importance weight from the text. The results of the study showed that the method was highly practical in different dataset types.

On the other hand, as a more mature algorithmic tool, the applications of long short-term memory (LSTM) are gradually diversifying. Yu et al. [12] explored the cellular learning capability of LSTM and distinguished them into two main parts: LSTM-dominated networks and integrated LSTM. The future directions of both algorithms were analyzed. Bukhari et al. [13] proposed a financial market forecasting model combining fractional order derivatives and LSTM neural networks. Results showed that the program was effective in financial data calculation. Alhussein et al. [14] combined networks with LSTM to form a comprehensive deep learning model, which could make a more comprehensive and accurate predictions of single-household short-term household electricity compliance. Ding et al. [15] proposed a prediction model combining LSTM and attention mechanism for the flood prediction problem, which applied variable control methods to the selection of hyperparameters and focused on the interpretation of attention weights. The results of the study showed that in most cases, the model was better than those of the same type of models and had some validity.

This study uses LSTM as the basic model, combines it with attention mechanism, and uses bi-directional encoder representations from transformers (BERT) model for text processing to form a higher performance long-text classification model, which has better application effects.

3. Design of Chinese Long-Text Classification Pattern

3.1 Long-Text Vector Processing Strategy

The proposed long-text classification model is broken into two portions. The first portion is a BERT training model for long-text word vectors and sentence vectors, and the second portion is a bi-directional long short-term memory (Bi-LSTM) model. The overall pattern tectonic diagram is denoted in Fig. 1.

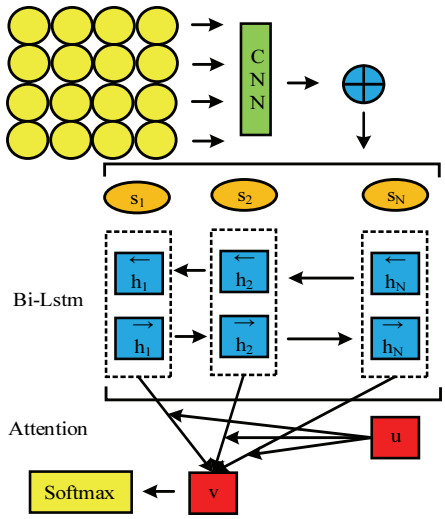

Fig. 1. Structure of long-text classification model.

The model structure includes five parts: word embedding, sentence vector integration, convolutional neural network (CNN), Bi-LSTM, and attention layer. The model first segments the long text, and then uses the BERT model to obtain sentence vectors and word vectors. Subsequently, the CNN layer captures local text features, combines sentence and feature vectors, and forms global text information through Bi-LSTM. The attention mechanism improves classification accuracy. In the model preprocessing step, images, chaotic characters, etc., are discarded, and only the text corpus is retained. After removing the stop words, the text length threshold is determined. According to the threshold, text is processed. If it is too long, it is deleted, and if it is too short, it is supplemented. Long text is processed using segmentation to meet BERT input requirements and avoid information loss. The model mainly includes input layer, convolution layer, pooling layer, and fusion layer.

The input layer is a matrix constructed based on word vectors in the form of $n \cdot d$, where n is the longitude representation of the local version and d is the dimensional representation of the word vectors. The single row of the input layer corresponds to the d-dimensional word vector corresponding to a single word, and the convolutional layer word is extracted using the multiple sets of convolutional kernels of different sizes to achieve the extraction.

$$c_i^j = f\left(WX_{i:i+h-1}^j + b\right). \tag{1}$$

In Eq. (1), W means the convolution kernel with the shape features $h \cdot d$. h is the height representation of kernel specification, d is the dimensional representation of word vector, and $X_{i:i+h-1}^j$ is the vocabulary

vector matrix from the *i*-th word to $i+1$ word of the j group of data in local text. $f(\cdot)$ activation function in the nonlinear state is represented by b, which is the quantized representation of the bias term. The convolution count relies on the vocabulary vector of the local text, and the result of the convolution operation can be obtained as Eq. (2):

$$c^j = \left[c_1^j, c_2^j, c_3^j, \ldots, \quad c_{n-h+1}^j\right]. \tag{2}$$

In Eq. (2), $n-h+1$ denotes the dimensionality specification of the result vector generated by the convolution kernel. The results obtained by the convolutional level operation need to be reduced by the pooling layer, and this dimensionality reduction process can help the model avoid the problem of overfitting. Then the stabilization of the patterns is improved in the process of part characteristic gain, and the study mainly adopts the maximum pooling method for the collection of the most text characteristic. The formula of maximum pooling operation is shown in Eq. (3):

$$\hat{c}^j = \max(c^j). \tag{3}$$

This study used different specifications of convolutional kernels in the feature extraction process, which are mainly divided into three different size specifications of small and large. The feature vectors formed by fusion after pooling of different size convolutional kernels are shown in Eq. (4):

$$z^j = \left[\hat{c}_1^j, \hat{c}_2^j, \ldots, \hat{c}_{3m}^j\right]. \tag{4}$$

The sentence vector formed by the BERT pattern is combined with the feature vector obtained by the CNN model, therefore final sentence vector of local text is formed as shown in Eq. (5):

$$s_j = l_j \oplus z^j. \tag{5}$$

In Equation (5), l_j denotes the BERT model forming sentence vectors, and z^j denotes the feature vectors obtained by the CNN model.

3.2 Bi-LSTM Model Design

The Bi-LSTM pattern used in the study mainly consists of a cell structure and three groups of gated cells, which are input-gated cells, forget-gated cells, and output-gated cells. The specific structures are shown in Fig. 2.

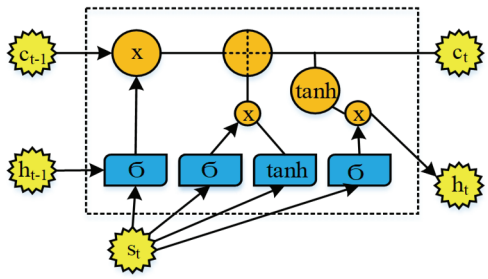

Fig. 2. Bi-LSTM model.

In Fig. 2, the Bi-LSTM model takes the part characteristic of text data s_t as the input vector based on the temporal order, c_{t-1} means recall unit of a moment ago of the moment, and h_{t-1} represents the result of a moment ago of the moment. Since the LSTM model used in the study is a bi-directional acquisition of the up and down two-way features of the text, the model forward hidden state output at the current moment is shown in Eq. (6):

$$\vec{h}_t = \overrightarrow{lstm}(s_t, \vec{h}_{t-1}). \tag{6}$$

The model backward hidden state output at the current moment is shown in Equation (7).

$$\overleftarrow{h}_t = \overleftarrow{lstm}(s_t, \overleftarrow{h}_{t+1}). \tag{7}$$

Since the model cannot focus on the data of all nodes, the study introduces an attention mechanism to assign weights or a weight to characteristics of the export data of different nodes for fusion. The specific structure is as Fig. 3.

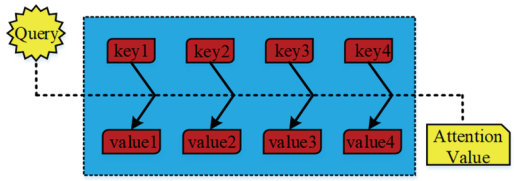

Fig. 3. Attention mechanism structure.

In Fig. 3, $Query$ represents the decoder, while key and $value$ are both encoders, and the decoder content is finally formed by the encoder to obtain the attention focus. In the process of attention calculation, the correlation degree of the encoder key and $value$ needs to be calculated first to get the weight coefficient corresponding to the encoder key. The specific calculation process of the weight coefficient is shown in Eq. (8):

$$\text{sim}_i = f(Query, key_i) \tag{8}$$

The $f(\ \ )$ in Eq. (8) denotes the scoring function of the attention mechanism. The final long-text classification probability is expressed as:

$$\hat{y} = soft\ \max(W_c v + b_c). \tag{9}$$

In Eq. (9), W_c denotes the weight matrix between the final hidden hierarchy of the model and the output classification, and b_c denotes the bias.

4. Analysis of the Application Effect of Long-Text Classification Model

4.1 Analysis of the Impact of Word Vectors and Local Text on Performance

When analyzing the application effect of the long-text classification model, the research will analyze

it from the local performance and the overall classification application effect. The training effects of models with different local-text lengths under the progression of training times are shown in Fig. 4.

From Fig. 4, although the overall performance convergence speed was not as fast as the advantages in maritime data collection environments, it still had the best performance among all specifications. It can be seen that L200 is the optimal local text length for this model.

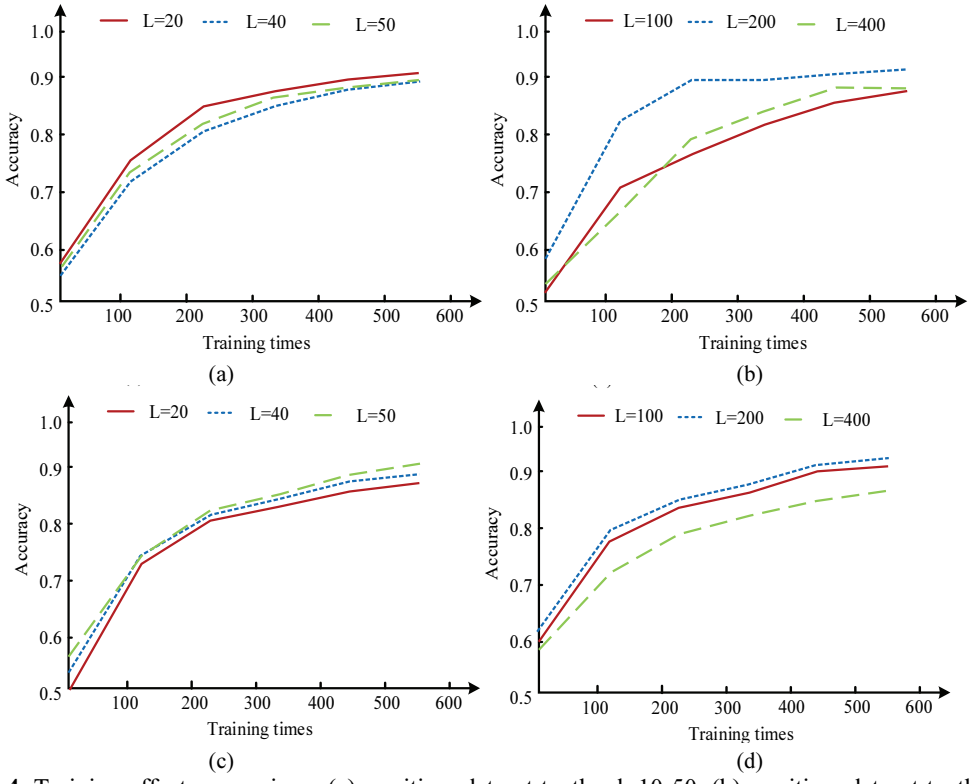

Fig. 4. Training effect comparison: (a) maritime dataset textbook 10-50, (b) maritime dataset textbook 100-400, (c) Chinese dataset of Fudan University textbook 10-50, and (d) Chinese dataset of Fudan University textbook 100-400.

4.2 Analysis of the Application Effect under Different Datasets

The study compared the model designed in the study with the recurrent neural network (RNN)-text model, CNN-LSTM model, and hierarchical attention networks (HAN) model, respectively. The comparative analysis of classification is in Fig. 5.

From Fig. 5, it can be seen that in the comparison of classification accuracy, in the maritime dataset environment, the classification accuracy of the model designed in this study was 92.93%. In the Chinese dataset of Fudan University, the classification accuracy of the model designed in this study was 93.77%, indicating that the classification accuracy of the designed model was the highest. In the comparison of calculation time, the model studied and designed in the maritime dataset had a calculation time of 85.42 seconds, while in the Chinese dataset of Fudan University, the running time of the model studied and designed was 117.51 seconds. It can be seen that the model studied and designed in the maritime dataset environment had the shortest calculation time.

5. Conclusion

To solve the problem of insufficient classification accuracy caused by the difficulty in extracting semantic features in long-text classification, this study improved the LSTM model and introduced an attention mechanism, incorporating the BERT model into the text vector processing classification model for long-text data. The accuracy of this model was 92.93% and 93.77%, respectively, which was the highest among the same type of models. The calculation time was 85.42 seconds and 117.51 seconds, respectively, which was the shortest among the same type of models. The research designed model had the fastest convergence speed among the same type, which meant that it achieved the best classification effect the fastest. It overcame the problem of classification failure caused by complex text features in long-text classification environments, and also provided the possibility for specific classification paths for long texts. Although the models designed in the study had significant advantages, research mainly focused on long-text classification. Therefore, designing comprehensive and robust long- and short-text classification models is the main research direction in the future.

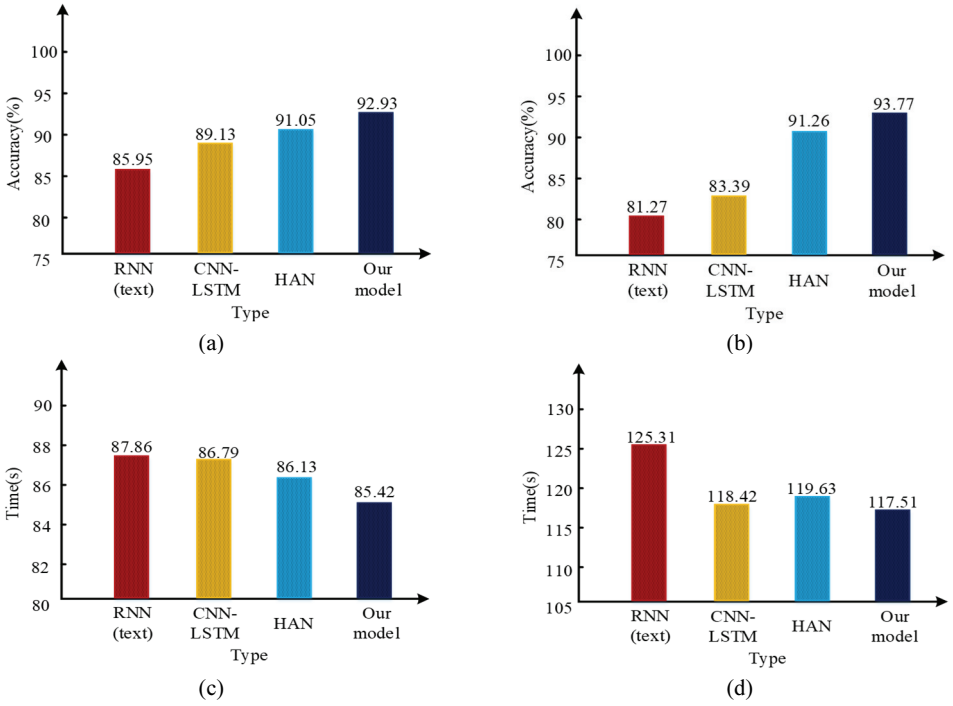

Fig. 5. Model comparison: (a) accuracy-maritime dataset, (b) accuracy-Fudan University Chinese dataset, (c) time-maritime dataset, and (d) time-Fudan University Chinese dataset.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

The research was supported by: Research on image super-resolution reconstruction based on depth learning (No. JJKH20221281KJ).

References

[1] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 273-292, 2019. https://doi.org/10.1007/s10462-018-09677-1

[2] C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, et al., "Transfer learning enables predictions in network biology," *Nature*, vol. 618, no. 7965, pp. 616-624, 2013. https://doi.org/10.1038/s41586-023-06139-9

[3] C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, Q. Ouyang, L. Lai, and J. Pei, "ransfer learning for drug discovery," *Journal of Medicinal Chemistry*, vol. 63, no. 16, pp. 8683-8694, 2020. https://doi.org/10.1021/acs.jmedchem.9b02147

[4] A. Hosseini, M. A. Eshraghi, T. Taami, H. Sadeghsalehi, Z. Hoseinzadeh, M. Ghaderzadeh, and M. Rafiee, "A mobile application based on efficient lightweight CNN model for classification of B-ALL cancer from non-cancerous cells: a design and implementation study," *Informatics in Medicine Unlocked*, vol. 39, article no. 101244, 2023. https://doi.org/10.1016/j.imu.2023.101244

[5] M. Ghaderzadeh, M. Aria, A. Hosseini, F. Asadi, D. Bashash, and H. Abolghasemi, "A fast and efficient CNN model for B-ALL diagnosis and its subtypes classification using peripheral blood smear images," *International Journal of Intelligent Systems*, vol. 37, no. 8, pp. 5113-5133, 2022. https://doi.org/10.1002/int.22753

[6] M. Lotfollahi, M. Naghipourfar, M. D. Luecken, M. Khajavi, M. Buttner, M. Wagenstetter, et al., "Mapping single-cell data to reference atlases by transfer learning," *Nature Biotechnology*, vol. 40, no. 1, pp. 121-130, 2022. https://doi.org/10.1038/s41587-021-01001-7

[7] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 7370-7377, 2019. https://doi.org/10.1609/aaai.v33i01.33017370

[8] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-level text classification using single-layer multisize filters convolutional neural network," *IEEE Access*, vol. 8, pp. 42689-42707, 2020. https://doi.org/10.1109/ACCESS.2020.2976744

[9] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325-338, 2019. https://doi.org/10.1016/j.neucom.2019.01.078

[10] H. Huan, Z. Guo, T. Cai, and Z. He, "A text classification method based on a convolutional and bidirectional long short-term memory model," *Connection Science*, vol. 34, no. 1, pp. 2108-2124, 2022. https://doi.org/10.1080/09540091.2022.2098926

[11] J. Zheng and L. Zheng, "A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification," *IEEE Access*, vol. 7, pp. 106673-106685, 2019. https://doi.org/10.1109/ACCESS.2019.2932619

[12] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235-1270, 2019. https://doi.org/10.1162/neco_a_01199

[13] A. H. Bukhari, M. A. Z. Raja, M. Sulaiman, S. Islam, M. Shoaib, and P. Kumam, "Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting," *IEEE Access*, vol. 8, pp. 71326-71338, 2020. https://doi.org/10.1109/ACCESS.2020.2985763

[14] M. Alhussein, K. Aurangzeb, and S. I. Haider, "Hybrid CNN-LSTM model for short-term individual household load forecasting," *IEEE Access*, vol. 8, pp. 180544-180557, 2020.

https://doi.org/10.1109/ACCESS.2020.3028281

[15] Y. Ding, Y. Zhu, J. Feng, P. Zhang, and Z. Cheng, "Interpretable spatio-temporal attention LSTM model for flood forecasting," *Neurocomputing*, vol. 403, pp. 348-359, 2020. https://doi.org/10.1016/j.neucom.2020.04.110

Li Lin **https://orcid.org/0000-0001-5953-9340**

She received a bachelor's degree in information and computing science from Jilin Agricultural University in 2016 and a master's degree from Northeast Normal University in 2019. She is currently a doctor of Jilin Agricultural University. Her research interests are data mining, text classification, agricultural machinery informatization and smart agriculture.

Yidan Wang **https://orcid.org/0000-0001-9241-139X**

She received a bachelor's degree in software engineering from Changchun University of Technology in 2017 and a master's degree from Changchun University of Technology in 2020. Currently, she is a doctoral candidate in intelligent mechanical and electrical equipment and control major of Changchun University of Technology. Her research interests include fault diagnosis, data analysis and intelligent processing.

Yan Wang **https://orcid.org/0000-0001-9578-1185**

He received a bachelor's degree in computer science and technology from Changchun University of Technology in 2003 and a master's degree in computer application technology from Changchun University of Technology in 2007. Currently, he is an associate professor of the School of Computer Science and Engineering, College of Humanities and Information Changchun University of Technology. The main research directions are software engineering and information security.

Changhua Tang **https://orcid.org/0000-0001-5668-8404**

He received a bachelor's degree in computer science and technology from Northeast Normal University in 2004 and a master's degree in computer science and technology from Northeast Normal University in 2007. Currently, he is an associate professor of the School of Computer Science and Engineering, College of Humanities and Information Changchun University of Technology. The main research directions are image processing and machine learning.