LOCAL SELF-ATTENTION BASED CONNECTIONIST TEMPORAL CLASSIFICATION FOR SPEECH RECOGNITION

Deng Huizhen and Zhang zhaogong

Computer Science Institution, Heilongjiang University of China, China

ABSTRACT

Connectionist temporal classification (CTC) has been successfully applied to end-to-end speech recognition tasks, but its main body recurrent neural network makes parallelization very difficult. Since the attention mechanism has shown very good performance on a series of tasks such as machine translation, handwriting synthesis, and image caption generation for loop sequence generators conditioned on input data. This paper applies the attention mechanism to CTC, and proposes a connectionist temporal classification based on the local self-attention mechanism, in which the cyclic neural network module in the traditional CTC model is replaced by the self-attention module. It shows that it is attractive and competitive in end-to-end speech recognition. The proposed mechanism is based on local self-attention, which uses a sliding mechanism to obtain acoustic features locally. This mechanism effectively models long-term scenarios by stacking multiple sliders to obtain a larger receiving field to achieve online decoding. Moreover, the CTC training joint cross-entropy criterion makes the model converge better. We have completed experiments on the AISHELL-1 dataset. The experiments show that the basic model has a lower character error rate than the existing state-of-the-art models, and the model after cross entropy has been further improved.

KEYWORDS

Connectionist temporal classification, self-attention mechanism, cross entropy, Speech Recognition

1. INTRODUCTION

End-to-end speech recognition is a recently proposed method, which does not require a pre-defined alignment between speech frames and characters to directly transcribe speech into text [1-9]. The latest work on end-to-end speech recognition can be divided into two main methods: based on connectionist temporal classification (CTC) [10,1-3] and attention-based encoder-decoder [4-6]. Both of these methods solve the problem of variable-length input and output sequences. In the traditional deep neural network hidden Markov model hybrid system, the deep neural network is used to generate each frame of sound data, and its distribution is re-expressed as the transmission probability of the Hidden Markov Model (HMM). Then, model training can be performed by using frame-level cross entropy (CE) criteria, using sequence discrimination training methods such as maximum mutual information (MMI) [8]. For this model, its problem is that the frame-level training target must be inferred from the alignment determined by the HMM. Different traditional speech recognition methods, the end-to-end model learns the mapping of acoustic frames to characters for the final target of interest, and tries to correct the sub-optimal problems caused by the irrelevant training process. Among them, the key idea of CTC is to use intermediate label representation, it allows duplicate labels and uses blank labels to identify

Dhinaharan Nagamalai et al. (Eds): CSEIT, WiMoNe, NCS, CIoT, CMLA, DMSE, NLPD - 2020
pp. 275-284, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101122

labels which are not output. The CTC loss can be effectively calculated by the forward and backward algorithm, which can predict the target of each frame, and provided that the conditions between the targets are independent of each other.

Recently, self-attention mechanism [11, 12] was proposed, which uses the entire sequence to model feature interactions at any distance in time. It is used in the encoder, decoder and feedforward context to accelerate the translation speed, and provides the latest translation results, sentiment analysis [13] and other tasks. The success of self-attention in these tasks inspired the initial work of self-attention in speech recognition. So the attention-based encoder decoder model appeared. Although it was first applied to machine translation, its versatility also made it useful for speech recognition tasks [14-17]. The attention-based encoder decoder model directly learns the mapping from the acoustic frame to the character sequence. At each output time step, the model sends out a label based on the history of the input and target labels. Since the attention model does not use any conditional independence assumptions, it exhibits a lower character error rate (CER) than CTC without using an external language model. However, some speech recognition tasks in real environments, the model shows poor results because the estimated alignment in the attention mechanism is easily damaged by noise and other details. Another problem is that it is difficult to learn the model from scratch due to the misalignment of long input sequences.

In order to overcome the above problems, this paper proposes a novel end-to-end speech recognition method, which uses a local self-attention model based on CTC training criteria to improve performance and accelerate learning. The key of our method is to use a shared encoder representation trained by CTC and self-attention model targets at the same time. We believe that the weakness of the attention model is due to the lack of left-to-right constraints used in DNN-HMM and CTC, which makes it is difficult to properly align the training encoder network under noisy data or long input sequences. Our proposed method improves the performance by correcting the CTC loss based on the forward and backward algorithm plus the cross-entropy loss function to assist the alignment problem of CTC training. In addition, combining the characteristics of attention and the defects of CTC, and inspired by time-delayed neural networks, this paper proposes a mechanism based on local self-attention that uses a sliding mechanism to obtain acoustic features locally, and stacked a larger receiving field to effectively model long-term scenarios to achieve online decoding.

2. RELATED WORKS

Recently, there have been some works applying the self-attention mechanism to speech recognition, and good results have been obtained compared with traditional hybrid speech models [18, 19]. Different from these, this paper introduces the self-attention mechanism into the CTC-based model and proposes a sliding mechanism similar to the convolutional neural network to achieve online decoding. Different from the block jumping mechanism in [19], this article divides the entire pronunciation into several overlapping blocks as input, and the slider has an asymmetric context. We use a sliding window at each layer to limit the scope of self-attention. They all use sliding windows to model the local dependencies between inputs, without any modification to the self-attention network structure. The sliding chunk mechanism only uses sliding windows to limit the range of attention, and stacks multiple self-attention sliders for long-term dependencies are modeled. Existing work [20] believes that only using CTC to train the model, sometimes the training failed to converge, or the cross-entropy loss function is used to pre-train the model, and then the CTC training model is used on this model, and there is still a problem of instability of the model. Therefore, this paper proposes to use CTC training and cross-entropy loss function at the same time to make the model converge better.

3. CONNECTIONIST TEMPORAL CLASSIFICATION

With CTC as the acoustic model sequence of the loss function, CTC can automatically learn the alignment between the input speech frame sequence and its label sequence (such as phonemes or characters) without using frame-level alignment information. Only one input sequence and one output sequence are needed for training. CTC cares about whether the predicted output sequence is close to the real sequence, and does not care whether each result in the predicted output sequence is exactly aligned with the input sequence at the time point. The CTC modeling unit is a phoneme or a word, and its main idea is to introduce Blank (-) tags, delete blank tags and merge duplicate tags to obtain a unique corresponding sequence. For a piece of speech, the last output of the CTC is a sequence of spikes, the position of the spike corresponds to the Label of the modeling unit, and the other positions are Blank.

For alphabet L, the size after adding the blank label (-) introduced by CTC is $L' = L \cup \{'-'\}$,Input an acoustic sequence x of length T, $x = (x_1,...,x_T)$, The corresponding output length is U tag sequence l, $l = (l_1,...,l_U)$, and U≤T.One way to match the input x neural network output $\pi = (\pi_1,...,\pi_T)$,It is defined as a sequence above L', which is $\pi \in L'^T$ .There is such a transformation function B, for all possible output paths after B transformed into label l, namely $l = B(\pi)$, This transformation removes the blank tags in the path and merges the consecutively repeated tags to obtain a unique corresponding sequence, for example, B(_,a,a,_b,_c,c,_)=abc,B(a,_,b,b ,_,_c,c,_)=abc. Therefore, for a given input x, the probability of its output sequence l after the neural network is the sum of the probabilities of all possible paths, the formula is as follows:

$$p(l \mid x) = \sum_{\pi \in B^{-1}(l)} p(\pi \mid x) \qquad (1)$$

which

$$p(\pi \mid x) = \prod_{t=1}^{T} y_{\pi}^{t} \qquad (2)$$

In order to facilitate the calculation of the gradient, the log objective function is generally minimized:

$$\ell_{ctc}(x) = -\log\ p(l \mid x) \qquad (3)$$

Since there are many possible paths and the amount of calculation is too large, CTC uses a forward-backward algorithm to calculate the loss, and uses a cluster search algorithm to decode.

4. MODEL

In order to improve the parallel computing power and performance of the model, this paper proposes an end-to-end model that does not use recurrent neural networks, a CTC model based on local self-attention, which uses a self-attention mechanism to replace the original CTC The recurrent neural network structure in the model. A slider mechanism similar to the convolutional neural network is proposed. The input acoustic feature length is 25% as the slider length and the features are stacked in chronological order. The experiment proves that this ratio has a certain effect.

4.1. Model Structure

For alphabet L, given an input sequence x of length T, $x = (x_1,..., x_T)$, x has T*d dimensions, defining an output sequence $y = (y_1,..., y_U)$ .In speech recognition, x is an acoustic feature, L is a collection of characters or phonemes, and the output sequence y is the corresponding real label on the alphabet. For the model structure of this article, the structure of the recurrent neural network replaced by the self-attention mechanism is shown in Figure 1, and the self-attention mechanism module is shown in Figure 2:

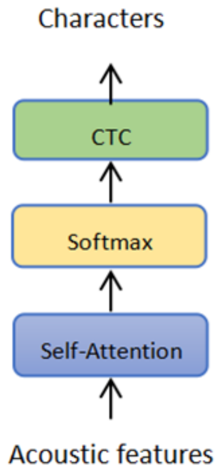

Figure 1. Basic model

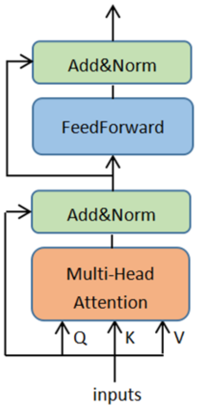

Figure 2.Self-attention block

This paper proposes a sliding mechanism similar to the convolutional neural network, which scans the acoustic features locally according to the 25% ratio of the input length as the slider length, as shown in Figure 3. The specific operation will be described in detail in section 4.1.2.

4.1.1. Multi-Head attention

Self-attention is a mechanism that associates different positions in the input sequence to calculate the input representation. Specifically, it has three inputs, namely query, key and value. The output of a query will be calculated as a weighted sum of values, where the weight of each value is calculated by the design function of the query and the corresponding key. Here, we use zoomed dot product attention, which is an effective self-attention mechanism, which has been demonstrated in [11]. As shown in Figure 2, Q represents the query, K is the key and V is the value, and the dimensions of the three variables are the input acoustic feature length multiplied by the model dimension, then the output of self-attention is:

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

The function of the scaling factor is to prevent the softmax function from entering an area with a very small gradient.On the basis of single-head attention, we adopt a multi-head attention mechanism, which calculates the dot product attention of h zooms, where h represents the number of heads. The original paper maps d_model (model dimensions) h times, each time three dimensions are obtained, d_q, d_k, d_v, and the attention value of each head is calculated in parallel, then the output of multi-head attention is:

$$MultiHead\,(Q,K,V) = Concat(head_1,...,head_h)W^O \tag{5}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{6}$$

The dimension of the mapping matrix W_i^Q is the model dimension multiplied by the query dimension, Similarly, the dimensions of W_i^k and W_i^V are the model dimension multiplied by the key dimension, the model dimension multiplied by the value dimension, and the query dimension equals the key dimension equals the value dimension. The dimension of W^O is the model dimension multiplied by the model dimension, and the above dimension is expressed by the formula:

$$d_q = d_k = d_v = d_{\mathrm{mod}\,el} / h$$

In addition, there is a position feedforward network layer after the multi-head attention layer, which contains two linear transformations and a RELU activation function:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{7}$$

The dimension of the mapping matrix W_1 is the model dimension multiplied by the feedforward network layer dimension, the dimension of W_2 is the feedforward network layer dimension multiplied by the model dimension, and the bias vector $b_1 b_2$ is learned, and the dimension is consistent with the model dimension. In these two After each sublayer, a layer normalization operation is connected, LayerNorm(x + Sublayer(x)), where Sublayer(x) is a function implemented by the sublayer itself.

4.1.2. Sliding Chunk mechanism

Since this article uses the self-attention mechanism, due to the characteristics of this mechanism, we must use the entire feature sequence as the input of the model to calculate the attention weight, and CTC originally did not achieve real-time output text, combined with the characteristics of attention and the shortcomings of CTC, and Inspired by the time-delay network, this paper proposes a local self-attention mechanism that uses a slider mechanism to obtain acoustic features locally. This mechanism effectively models long-term models by stacking multiple sliders to obtain a larger receiving field. Scenario to achieve online decoding function. In addition, regarding the size design of the slider, according to the input acoustic feature size, a fixed-length slider is taken by multiplying the acoustic feature length by a ratio of 25%. As shown in Figure 3, the fixed-length slider flows along the time axis of the feature sequence. Moreover, stacking multiple self-attention blocks makes it possible to model a longer time context without causing excessive performance degradation.

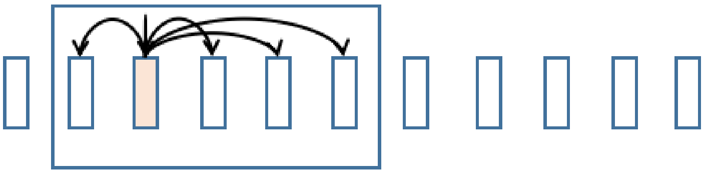

Figure 3. Local self-Attention block

The attention module maps the input x_t into three vectors: $q_t, k_t v_t$ represents the query, key and value, respectively, and the output h_t is the weighted sum of the value v_t (varying with time), where the weight is determined by the dot product of the query and the key (through the softmax activation function Standardization) decision. For the single head example, it can be defined by the following formula:

$$h_t = \sum_{\tau=t-L}^{t+R} c_{t,\tau} v_t \tag{8}$$

Where ct(τ) = exp(qt · kτ)/Zt, Zt represents the normalization operation, which guarantees $\sum_{\tau} c_{t,\tau} = 1$. In a fixed-length slider, L and R represent the frames to the left and right of the current time t, respectively number. For the long example, its formula is:

$$h_{i,t} = \sum_{\tau=t-L}^{t+R} \alpha_{i,\tau} s_\tau \tag{9}$$

Where $\alpha_{i,\tau} = Attention(s_\tau, K, V)$ and K, V are the τ-th vector in the slider, hi, t represents the i-th head in the multi-head attention layer at time t, and s_τ, K, V represent the τ vectors in the slider. The slider length of each block is L+R+1.

4.2. Loss Function

Existing work believes that only the CTC training model is used, and sometimes the training fails to converge, or the cross-entropy loss function is used to pre-train the model. If the CTC training model is used on this model, there is still the problem of instability and model instability. Therefore, this paper proposes to use CTC and cross entropy at the same time to make the model converge better. Then the loss function after using CTC and CE jointly is as follows:

$$\ell_{joint}(x) = \ell_{ctc}(x) + \ell_{ce}(x) \tag{10}$$

$$\ell_{ce}(x) = -\sum_{i=2}^{K}(1 - p(y_1 \mid x))t_i \log p(y_i \mid x) \tag{11}$$

The CTC loss function is shown in formula (3), $p(y_1| x)$represents the probability of the CTC blank label of the softmax output layer, and the probability of 1 minus the blank label is used as the normalization factor of the cross-entropy loss function, t_i (i=2, ..., K) represents the target label at the frame level. This normalization factor plays an important role. At the beginning of training, the prediction of the acoustic model is like random guessing, and then CTC and cross-entropy loss function both play an important role in guiding the training. In the training process, the CTC loss often produces a shape peak distribution, each output target has only a few peaks, and the rest of the time is likely to predict blank labels. Therefore, the standardized cross-entropy loss function will help produce accurate alignment for the output target without affecting the allocation of blank labels. As a result, the proposed joint CTC-CE training will be more stable and help alleviate the delay problem.

5. EXPERIMENT

5.1. Dataset

The experiment mainly performed speech recognition tasks on Chinese (Mandarin). The open source speech corpus AISHELL-1 [21] is used for Mandarin speech recognition, and all speech files are sampled at 16 K Hz and 16 bits. The training set contains 150 hours of speech recorded by 340 speakers; the development set contains 20 hours of speech recorded by 40 speakers; and the test set contains 10 hours of speech recorded by 20 speakers. And the speakers in the training set, development set and test set do not overlap.

5.2. Experiment Set

As mentioned earlier, this article uses a connectionist-based temporal classification (CTC) speech recognition system. When doing the experiment, we used the 40-dimensional Mel filter library coefficient feature calculated on a 25ms window with a 10ms displacement. Each feature is rescaled so that the mean and unit variance of each audio sample is zero. When the processed frame is at time t, the number of left and right frames is asymmetrical, and these features are stacked in time through the local self-attention module, and finally down-sampled to a 30ms frame rate. This paper selects 4231 characters (including "blank" characters) as the model unit. During training, all audio samples are sorted by length and modeled using PyTorch [22], and Kaldi [23] is used for data preparation. Like the attention mechanism [2], this paper uses 6 self-attention models as encoders, in which the dimensions of query, key and value are all 128, the number of attention heads is 8, the model dimension and feedforward network layer The dimension of is 1024. Due to the introduction of local self-attention (see section 4.1.2), this article does not use the position coding formula in the original text of attention to reduce the amount of calculation. After that, an initial learning rate of 0.001 is used to train the network to reduce the joint loss function of CTC and CE.

5.3. Results and Discussion

Table 1. The CERs (%) of the development and test sets of AISHELL-1.

Model	Dev.	Test.
Baseline	7.36	8.51
Baseline+CE	7.29	8.42
BN	8.35	9.71
ABN-U	7.40	8.40

As shown in Table 1, we compared the character error rates of the four models on the AISHELL-1 dataset. Dev. and Test. respectively represent development dataset and test dataset.The Baseline model represents the model proposed in this article, and the second is the joint training with cross entropy. Model, BN is the basic model in [24], it is an acoustic model based on cyclic neural network and using CTC training, ABN-U is the sentence-level attention batch normalization model in [24]. From the data in the table, it can be seen that on the development dataset, the model proposed in this article is relatively better. On the test set, the model performs better than the basic model after adding cross entropy, and the effect is similar to ABN-U.

6. CONCLUSION

In this work, we propose a local self-attention encoder, which replaces the recurrent neural network with a self-attention module. The performance of the self-attention encoder is better than the BN (CTC original model under the same data set) model. Use local self-attention mechanism (slider mechanism) to realize online decoding function. Moreover, this paper also proposes a CTC joint cross-entropy criterion training model to improve model stability and facilitate better convergence of the model. The results show that the CTC joint cross-entropy criterion training method has greatly improved the basic model. During decoding, we observed that the model can predict characters with similar pronunciations well. In future work, we will explore how to optimize the parameter estimation problem of language models and unknown distribution data.

REFERENCES

[1] Alex Graves and Navdeep Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in Proceedings of the 31st International Conference on Machine Learning(ICML-14), 2014, pp. 1764–1772.

[2] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al., “Deep speech: Scaling up endto-end speech recognition,” arXiv preprint arXiv:1412.5567,2014.

[3] Yajie Miao, Mohammad Gowayyed, and Florian Metze,“EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding,” in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU).IEEE, 2015, pp. 167–174.

[4] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “End-to-end continuous speech recognition using attention-based recurrent NN: First results,” arXiv preprint arXiv:1412.1602, 2014.

[5] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk,Kyunghyun Cho, and Yoshua Bengio, “Attention-based models for speech recognition,” in Advances in Neural Information Processing Systems, 2015, pp. 577–585.

[6] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals,“Listen, attend and spell,” arXiv preprint arXiv:1508.01211,2015.

[7] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, “End-to-end attentionbased large vocabulary speech recognition,” arXiv preprint arXiv:1508.04395, 2015.

[8] Liang Lu, Xingxing Zhang, and Steve Renals, “On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5060–5064.

[9] William Chan and Ian Lane, “On online attention-based speech recognition and joint mandarin character-pinyin training,” Interspeech 2016, pp. 3404–3408, 2016.

[10] A. Graves, A. Fernandez, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the 23rd International Conference on Machine Learning. IEEE, 2006.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

[12] J. Cheng, L. Dong, and M. Lapata, “Long short-term memorynetworks for machine reading,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2016, pp. 551–561.

[13] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang,“DiSAN: Directional self-attention network for RNN/CNNfree language understanding,” in Proc. AAAI Conf. Artificial Intell. (AAAI), 2018.

[14] Y. Zhang, W. Chan, and N. Jaitly, “Very deep convolutional networks for end-to-end speech recognition,” in Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP). IEEE,2017, pp. 4845–4849.

[15] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in Proc. IEEE Int. Conf. Acoustics Speech Signal Process.(ICASSP). IEEE, 2017, pp. 4835–4839.

[16] C.-C. Chiu, T.N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen,Z. Chen, A. Kannan, R.J. Weiss, K. Rao, K. Gonina,et al., “State-of-the-art speech recognition with sequence-tosequence models,” in Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP). IEEE, 2018, pp. 4774–4778.

[17] A. Zeyer, K. Irie, R. Schluter, and H. Ney, “Improved training of end-to-end attention models for speech recognition,” in Proc. Ann. Conf. Int. Speech Communication Assoc. (INTERSPEECH), 2018.

[18] L. Dong, S. Xu, and B. Xu, “Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5884–5888.

[19] L. Dong, F. Wang, and B. Xu, “Self-attention aligner: A latencycontrol end-to-end model for asr using self-attention network and chunk-hopping.” arXiv: Computation and Language, 2019.

[20] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, “Learning acoustic frame labeling for speech recognition with recurrent neural networks,” in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015, pp. 4280–4284.

[21] H.Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in O-COCOSDA2017.

[22] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., “The kaldi speech recognition toolkit,” IEEE Signal Processing Society, Tech. Rep., 2011.

[24] Fenglin Ding, Wu Guo, Lirong Dai, Jun Du,“Attentive batch normalization for lstm-based acoustic modeling of speech recognition，″ CoRR abs/2001.00129， 2020

AUTHORS

Huizhen Deng was born in Anhui, China, in 1995.She received the B.S. degree in Qindao College of Qingdao Technological University, China, in 2018.She is currently pursuing the M.S. degree in computer science and technology in Heilongjiang University.

Zhaogong Zhang, professor, Ph.D, postdoctoral, master supervisor, Department of Software and Theory, School of Computer Science and Technology, Heilongjiang University. Member of the Big Data Division of Heilongjiang Institute of Industry and Application. He received a bachelor's degree in basic mathematics, a master's degree in basic mathematics and a doctorate degree in computer software and theory from Harbin Institute of Technology in the Department of Mathematics of Hei longjiang University in 1985, 1991 and 2003, respectively. His main research interests include bioinformatics, data mining, statistical genetics, big data, cloud computing, etc.