COMBINED WITH THE RESIDUAL AND MULTI-SCALE METHOD FOR CHINESE THERMAL POWER SYSTEM RECORD TEXT RECOGNITION

Jun LIU [1,2], Wei LI [1,2 *], Zhuang DU [1,2]

[*1] Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan, China

[2] School of Computer Science & Engineering, Wuhan Institute of Technology, Wuhan, China

[*] Wei LI; E-mail: xiaoaowen@wit.edu.cn

Abstract:Aiming at the problem that the recognition accuracy based on convolutional neural network of thermal power system record text is not high, a method of thermal power system record text recognition based on residual and multi-scale feature combination was proposed and implemented. Combined with the residual, a new network is designed to replace the traditional convolutional neural network and improve the feature extraction ability of the network. The 1×1 convolution core was used to increase the network depth and reduce the parameters instead of the 3×3 convolution core. the network order of each layer in residual block was adjusted so that the network representation ability can be improved. Combining feature information of different scales and retaining more vertical feature information, the classification accuracy of the network is improved. Experiments on the self-built image data set of thermal power system records show that the proposed network model improves the accuracy by 11% compared with CRNN, and has better robustness to image distortion and blurring.

Key words*: residual network; text recognition; thermal power system record text; convolutional neural network*

1. Introduction

At present, text recognition technology [1] is more and more widely used in the thermal power system field. With the help of the text recognition technology[2], thermal power system record information can be directly transcribed into a database for storage, which can greatly save labor and time costs. For the thermal power system record text image, it is different from the general scanned document that is clear and neatly arranged. However, it will also be affected by the same objective factors as the natural scene text recognition, such as illumination and tilt; Some old thermal power system records may be affected by objective factors such as time and paper, resulting in more complex background of actual thermal power system records, and even the content is missing. Most of the texts in the thermal power system record are written by Chinese. Compared with English recognition, it is necessary to overcome various problems such as various types of Chinese, complicated structures, diverse combinations and similar characters. for non-Chinese characters, such as the English abbreviation of thermal power system terminology and various special symbols, there will be a random combination with Chinese, and the mixing of different languages will increase the difficulty of recognition.

The traditional character recognition technology is mainly based on the structural feature information[3-4] or statistical features of characters[5]. Although the structural features can achieve high matching degree and effectively distinguish the near-shape characters, but they are extremely vulnerable to interference from external conditions and have poor stability. Through statistical features, such as grid features and directional pixel features, its anti-interference ability is enhanced, but it will reduce the ability to recognize the details of characters. Both structural features and statistical features are artificially designed, which not only consumes time and energy, but also may introduce some unnecessary noise that will lead to the accumulation of errors. These features are usually not universal and, in many cases, cannot even be effectively extracted, such as stroke features, shape features, edge features and so on.

With the rapid development of deep learning, the traditional text recognition technology has been gradually surpassed by the recognition technology with convolution neural network (CNN) as the core[6-16]. Deep learning provides a good solution to the problem of low recognition accuracy caused by extremely complex text background, abundant text types, random distribution, difficult character segmentation, noise and other factors. Through "learning", the deep learning model can effectively extract more abstract feature information in the image, which greatly improves the recognition accuracy, and has better anti-interference ability. At present, text recognition methods based on deep learning can be classified into character-level recognition and text sequence recognition. For CNN, the input and output dimensions of the data are fixed, Generally, a fully connected layer is connected as a classifier after CNN[17]. Other classifiers can also be used, For example, Yu [18] et al. combined CNN and support vector machine (SVM), which replaces the final output layer with SVM on the basis of lenet-5 [19] convolutional neural network, and uses CNN to extract feature information, and SVM is used for final classification. In practice, words usually appear continuously, and the network based on character recognition can only give one recognition result at a time, which requires character segmentation [20]. However, in complex cases, character segmentation is very difficult, and forced segmentation may destroy the character structure and the semantic relationship between characters. In order to realize text sequence recognition, it is necessary to learn the cyclic neural network in depth. CRNN[21], which combines the characteristics of CNN and RNN is the classical model of text sequence recognition. The sequence feature information of text is obtained by RNN, and variable length text recognition is realized by CTC. At present, most of the mainstream sequence recognition models adopt the design pattern of CRNN.

2. Related theory

2.1. Residual network

For convolution neural network, improving the network depth is an effective method to improve the network performance, but one problem with increasing network depth is that these additional layers are signals of parameter updates, because the gradient is propagating backwards and forwards, when the network depth is increased, the gradient of the upper layer will be very small, resulting in the stagnation of the network's learning of these layers, that is, the disappearance of the gradient. With the deepening of network depth, the problem of network training hard also highlighted, then there is the problem of network degradation, as the network gets deeper, the number of parameters increases, the large parameter space makes the optimization problem more complex, and simply increasing the

network depth will lead to the increase of training error, although the network converges, there is a degradation problem. The residual network ResNet proposed by He [22] et al. solves this problem well, and deeper network can be realized through the residual module.

The data in each residual module is divided into two parts: a regular route, and a shortcut, which is similar to a "short circuit" in a circuit, where shortcut's value is the input value of the layer's residual module, the main line is the common combination of convolutional neural network, or layer of normalized Batch Normalization (BN)[23], the activation function (relu) layer and Convolution Convolution. Its structure is shown as follows:

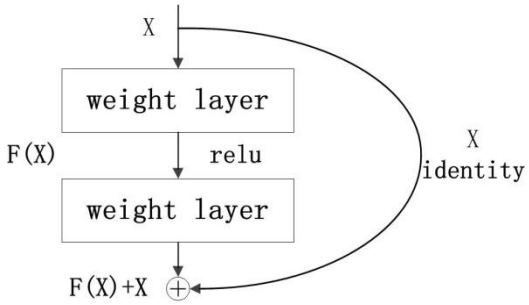

Fig.1 Residual module structure

Through the residual module, the input-output relationship of the general neural network can be expressed as $y = H(x)$, with the deepening of the network, the gradient directly calculated will encounter the degradation problem, if a residual module with shortcut structure has been used, then the goal of variable parameter optimization is not just $F(x)$, but to $F(x) = H(x) - x$, F(x) can be understood as a disturbance relative to the original input, which makes the network more sensitive to changes in output and easier to learn from than it was before.

2.2. LSTM

RNN can deal with the problem of long-term dependence in theory, but in practical application, it is found that with the increase of time span and the deepening of network, RNN can't deal with the problem of long-term dependence, it cannot get the long distance historical information, and it is prone to the problem of gradient disappearance. The LSTM proposed by Hochreiter [24] has well solved this problem. Its network structure is as follows:

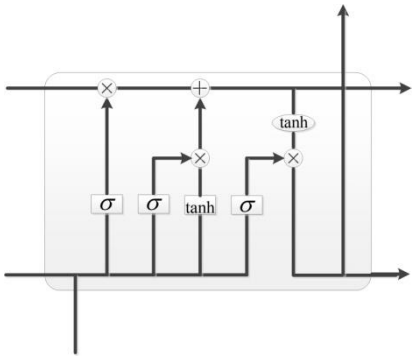

Fig.2 LSTM unit internal structure diagram

LSTM is a kind of repeated neural network chain form of modules like RNN, each step of the traditional RNN hidden unit just perform a simple activation with tanh or relu, but each cell of the LSTM is composed of three parts in the above, from left to right: forget gate, input gate and output gate, by using the gate mechanism, LSTM can selectively memorize information, so that the network

can maintain the ability of long-distance dependence on information, while reducing the problem of gradient disappearance

The forget gate decides to discard some information from the cell state. The gate reads h_{t-1} and x_t.The gate connects a sigmoid activation function to output a number from 0 to 1, and passes it to the C_{t-1} representing the cell state, the closer the number to 1, the more information is reserved, close to 0 means that the more information is to be left out. The update formula of forget gate data is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

In the above equation (1), f_t represents the output of forget gate, W_f is the weight, b_f is the bias and σ is the sigmoid activation function. The input gate contains two branches, one is the connection activation function sigmoid to determine the information i_t to be updated, the other is the connection activation function tanh to generate $\widetilde{C_t}$ to be updated, and finally, combined with the output of the forgotten gate, the state of the cell is updated, that is, C_{t-1} is updated to C_t. The formula for updating the output gate data is as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$C_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (4)$$

Finally the output of the gate also contains two branches, one based on the cell of the input data h_{t-1} and x_t will be determined by a sigmoid output information o_t, another branch based on the updated cell state C_t, connect tanh activation function to output a number between -1 to 1, which multiply with o_t to get the output h_t of the current time step of the network.The formula for updating the output gate data is as follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

2.3. Loss function

The Connectionist Temporal Classification (CTC)[25] is an algorithm for solving the classification problem of time series data. The CTC algorithm does not require the input and output to be strictly aligned. The result of the alignment is output by introducing a new placeholder. This placeholder is called a blank placeholder, which is represented by the symbol ε . The CTC contains three steps:

1. Combine consecutively repeated characters, including blank placeholders;
2. Remove all blank placeholders;
3. Connect the remaining characters.

The premise of CTC is that the alignment of the input and output is monotonous, the input and output are in a many-to-one relationship, and the length of the output is also smaller than the input.

The task of the CTC is to calculate the probability of all possible tag sequences given the network input. During training, the CTC needs to calculate the set of tag sequences with the highest probability and then learn to make the correct tag sequence. The probability is maximized. Assuming the input is X and the output is Y, the goal of the CTC is to maximize the probability of the following:

$$p(Y|X) = \sum_{\alpha(\pi)=Y} \prod_{t=1}^{T} p(a^t|X) \quad (7)$$

For the model of LSTM+CTC, the above equation shows the probability that the output class of LSTM network is a when the time step is t, T represents the sum of all-time steps of the LSTM, and the multiplication represents a path that can form a path length of which is T, and αrepresents the

merge operation of CTC.,which makes T mapped to the target sequence Y. Assuming that the sequence set of T consisting of a label set and a blank placeholder is L^T, and the set of labels without a blank placeholder length of which is less than the sequence T is $L^{T'}$,Then there are $\pi \in L^T$ and $Y \in L^{T'}$. Since the result in Equation 7 needs to be minimized, the CTC loss function can be obtained by negative log likelihood (D stands for training set):

$$\text{loss} = -\sum_{(X,Y\in D)} \ln(p(Y|X)) \qquad (8)$$

3. Implementation

Chinese, English and other special characters are included in the thermal power system record text.The characters are various and the Chinese structure is relatively more complicated. The thermal power system record text also has some characteristics of the scene text. Due to the influence of objective factors, the image quality is generally poor, which further improves the difficulty of identifying the thermal power system record text, Simple convolutional neural network has been unable to extract effective feature information to achieve accurate character classification. In order to improve the network representation ability, the text combines the characteristics of residual network, improves the network depth, and at the same time combines multi-scale feature information. The overall structure of the network is shown in the following figure:

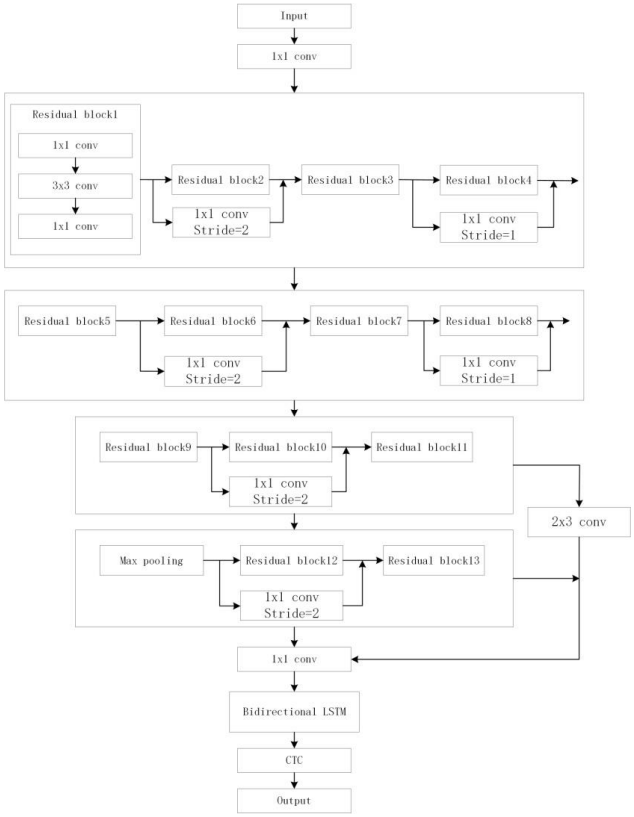

Fig.3 Network structure

Compared to the use of a 7-layer convolutional network in CRNN for image feature information extraction, the text proposes a residual network with 47 layers of convolutional layers. The first and last convolutions use a 1×1 convolution kernel to change the number of channels in the image, using the traditional Conv+BN+Relu structure. The result of the first layer convolution is input to the subsequent residual module, and each residual module contains 3 layers of convolution. To reduce the

parameter amount, the first layer and the third layer in the residual module are set to 1*1 Convolution, the second layer uses 3×3 convolution, and the data dimension changes in the residual module are divided into two categories: 1. The feature map size will change , 2. The channel number and the feature map size change will change, using the 1 or 2 step size or 1×1 convolutional layer implementation respectively. When the Residual branch in the residual module adopts the traditional Conv+BN+Relu structure, due to the characteristics of the Relu activation function, the output value of the branch is always a non-negative value, and the input will monotonically increase for forward propagation, affecting the network. The representation ability of the Residual branch uses the BN+Relu+Conv structure to ensure that the output of the branch is in progress. In order to enable the network to utilize multi-level feature information, the text combines the output of the eleventh residual block, the thirteenth residual block, and the last layer of the 1×1 convolution layer, and the dimensional information outputted by the network is batch, Channels, height, and width, where the 11th residual block corresponds to the output dimension (64, 128, 5, 33), the 13th residual block output dimension is (64, 256, 2, 33), and the last layer of the convolutional layer output dimension For (64, 512, 2, 33), since the subsequent access to LSTM is required for sequence feature extraction, simple splicing will destroy the spatial information of the feature. Therefore, this paper combines the feature maps of three different scales in the channels dimension, in the 11th After the difference block, connect a 2×2 convolutional layer and convert the output to (64, 128, 4,33), then convert to (64, 256, 2, 33) and the final merged data dimension is (64, 1024, 2, 33).

After the process of residual network are two bidirectional Long-Directional LSTM network layers. Each layer is composed of a forward LSTM and a backward LSTM, which can better capture the sequence feature information of the text. For the final output of the residual network, this paper divides it into 33 time steps in the width dimension. Since it is a bidirectional LSTM, the input data of each time step is a 2048-dimensional vector, and each LSTM network contains 256 cells. The first layer of bidirectional LSTM will output a 256-dimensional vector for each time step, and the second layer of bidirectional LSTM will output a vector dimension of 7067 for each time step, which is the sum of the total number of characters to be classified and the blank placeholder. Finally, the probability of the possible text sequence is calculated by combining the CTC, and the sequence with the highest probability is selected as the result output.

4. Experiment and analysis

4.1. Experimental environment

The experiment is based on the pytorch framework. Hardware environment:

- Intel(R) Core(TM) i7-8700 CPU @ 3.20 GHz
- 16G 4DDR memery
- GIGABYTE AORUS GeForce® GTX 1080 Ti Xtreme Edition 11G

4.2. Experimental data

Since there is no public data set related to thermal power system record text, the data set used in this paper mainly consists of synthetic data and artificial intercept data. Actual thermal power system record text is aging or incomplete， the text in the image is usually blurred, tilted, broken or stuck, and the background is usually more complicated than the general scanned text image, Therefore, this paper has carried out various processing on the synthesized image:

1. Rotate to make the text in the image tilt;
2. Use median blur and Gaussian blur to smooth;
3. Distort the text in the image;
4. Add Gaussian noise, salt and pepper noise, uniform noise, Rayleigh noise or gamma noise to the image;
5. Perform affine transformation and elastic transformation on the image;
6. Corrosion and expansion.

All of the above operations are randomly added to the image to ensure the diversity of the synthesized image.

4.3. Experimental results and analysis

In order to verify the performance of the proposed method, this paper conducted multiple sets of comparative experiments from different angles. For different method, this article uses the Adam [26] optimizer, the Batch is set to 64, and the maximum number of training iterations is 200.

The method proposed in this paper is compared with several other methods. Resnet-1 is based on the recognition method of traditional residual network. Resnet-2 improves the residual block based on Resnet-1， and the test set contains 56,868 Chinese thermal power system record text images. The results are shown in Table 1.

Table 1 Accuracte rate of different methods

methods	accurate rate
CRNN	83.63%
Resnet-1	90.84%
Resnet-2	92.15%
proposed method	95.07%

Compared with CRNN, the recognition modesl based on the residual network has higher recognition accuracy. For actual Chinese thermal power system record text images, The recognition accurate rate of CRNN is 83.63%. Obviously, Resnet-1 has higher accurate rate than CRNN, which is increased by about 7%. This shows that the residual network can extract the features of the text more effectively than the convolutional neural network. The difference between Resnet-2 and Resnet-1 is that the relative positions of the layers in the residual block are different. It can be found that by ensuring that the output range of the Residual branch of the residual block is $(-\infty, +\infty)$, and the data distribution of the Identity branch does not change, the accurate rate of Resnet-2 are increased by about 1% compared with Resnet-1.

The proposed method uses the residual network to construct the feature extraction network and combines the feature information of different scales to further improve the representation ability of the network. It can be seen from the table 1 that after training the above four models in the same way, and ensuring that the training data set and the test data set are the same, the method proposed in this paper proposed method achieves the highest accurate rate, reaching 95.07%, compared with the CRNN network recognition accuracy has been significantly improved.

In order to further verify the effectiveness of the proposed method, this paper compares the convergence speed with the CRNN. The convergence results of the two on actual Chinese thermal power system record text image data set are shown in the figure below.

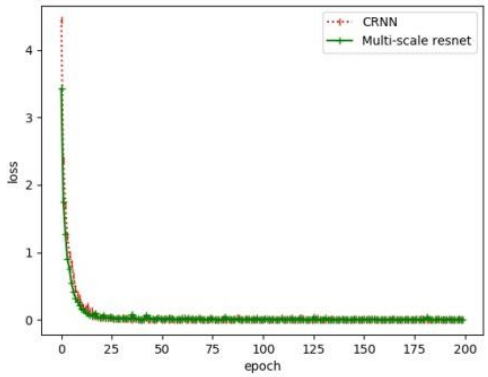

Fig.4 Convergence result

As the number of iterations increases, both models gradually converge, but it can be seen that the model proposed in this paper has a faster convergence speed.

The final output feature map of the model proposed in this paper is a one-dimensional vector, that is, the longitudinal feature information of the image has only one value. In order to explore the influence of the number of vertical feature information on the recognition rate, the text selects different vertical feature information dimensions for comparison. The results are shown in Table 2.

Table 1 Accuracte rate of different vertical feature dimensions

Vertical data dimension	accurate rate
1	93.91%
2	95.07%

It is found from the above table that advanced feature information is not necessarily more effective than low-level features. By adjusting the final vertical output dimension, this paper does not use the method of compressing images into one-dimensional vectors in the general sequence recognition algorithm, but add the vertical feature information dimension and the input data dimension of the bidirectional LSTM. When the vertical dimension is 2, the recognition accuracy is increased by about 1% compared to the vertical dimension of 1.

In the actual Chinese thermal power system record text recognition, the accuracy of the text detection algorithm will directly affect the accuracy of the recognition. Due to the poor quality of the Chinese thermal power system record image, a certain part of the character may be missing due to tilt or detection algorithm. Generally, it is concentrated on the front and rear ends of the text image. For this purpose, some Chinese thermal power system record text images containing characters are collected for testing. The test structure of different methods is shown in Table 3.

Table 1 Accuracte rate of incomplete character

methods	accurate rate
CRNN	37.11%
Resnet-1	68.04%
Resnet-2	72.16%
proposed method	74.23%

It can be seen from the table that the recognition method based on the residual network is significantly higher than the CRNN in the case of poor quality of test samples or containing some incomplete characters. By combining multi-scale feature information, and retaining more vertical

feature information, The proposed method still has the highest accurate rate, which proves that the proposed method has better robustness.

5. Conclusion

Combining the residual network and multi-scale feature information, this paper proposes a Chinese thermal power system record text recognition method. Firstly, the residual network are used to ensure that the network does not degenerate when the network is deeper. Then the internal structure of the residual block is optimized, which can extract the feature information of Chinese thermal power system record text image more effectively. Finally, the method combines the feature information of different scales and retains more vertical feature information. Compared with CRNN and traditional residual network, the proposed method improves the recognition accuracy effectively. At the same time, for the samples with poor quality and incomplete characters, this method also achieves the highest recognition accuracy and shows a good robustness

Acknowledgements

This work was supported by the National Natural Science Foundation of China（61172150，61803286）、the Foundation of Hubei Provincial Key Laboratory of Intelligent Robot (HBIR 201802) and the tenth Graduate Innovation Fund of Wuhan Institute of Technology(CX2018197, CX2018200, CX2018212).

References

[1]Mori S, *et al*., Historical review of OCR research and development. *Proceedings of the IEEE*, 80(1992), 7, pp. 1029–1058.

[2]Zhang Y, *et al*., Scene text recognition with deeper convolutional neural networks[C]// *IEEE International Conference on Image Processing*. 2015.

[3]Yao C, *et al*., Strokelets: A Learned Multi-scale Representation for Scene Text Recognition[C]. *Computer Vision and Pattern Recognition*, 4042-4049.2008, (2014), 11, pp. 134–135.

[4] Heutte L, *et al*., A structural/statistical feature based vector for handwritten character recognition[J]. *Pattern Recognition Letters*, 19 (1998), 7, pp. 629-641.

[5] Das N, *et al*., A statistical-topological feature combination for recognition of handwritten numerals[J]. *Applied Soft Computing*, 12 (2012), 8, pp. 2486-2495.

[6] YunTao W, *et al*., End-to-End Text Recognition with Convolutional Neural Networks[C]//*International Conference on Pattern Recognition*. 2013.

[7] Jaderberg M, *et al*., Reading Text in the Wild with Convolutional Neural Networks[J]. *International Journal of Computer Vision*, 116 (2016), 1, pp. 1-20.

[8]Ahranjany S S, *et al*., A very high accuracy handwritten character recognition system for Farsi/Arabic digits using Convolutional Neural Networks[C]. *bio-inspired computing: theories and applications*,（2010）， pp. 1585-1592.

[9] Wang Z, et al., Trilateral constrained sparse representation for Kinect depth hole filling. *Pattern Recognit Lett* 65 (2015), pp.95–102

[10] Liu H, et al., Distributed source localization under anchor position uncertainty. *Chin J Electron* 23 (2014), 1, pp. 93–96
[11] Zhong L, et al., OHRank: an algorithm integrating mentality and influence of opinion holder for opinion mining. *Chin J Electron* 22 (2013), 4, pp. 655–660
[12] Peng, Li , et al. A robust method for estimating image geometry with local structure constraint. *IEEE Access* 99(2018), PP. 1-10.
[13] Lu, Tao , et al. Robust Face Super-Resolution via Locality-constrained Low-rank Representation. *IEEE Access* (2017), pp. 1-10.
[14] Wang Z, *et al*., Trilateral constrained sparse representation for Kinect depth hole filling[J]. *Pattern Recognition Letters*, 65 (2015), pp. 95-102
[15] YunTao W, *et al*., Utilizing Principal Singular Vectors for 2D DOA Estimation in Single Snapshot Case with Uniform Rectangular Array[J]. *INTERNATIONAL JOURNAL OF ANTENNAS AND PROPAGATION*,(2015)
[16] YunTao W, *et al*., HOSVD-Based Subspace Algorithm for Multidimensional Frequency Estimation Without Pairing Parameters, *CHINESE JOURNAL OF ELECTRONICS*, 23 (2014), 4, pp. 729-734
[17]Jiao L, *et al*., Offline handwritten English character recognition based on convolutional neural network[C]// *Iapr International Workshop on Document Analysis Systems*. 2012.
[18]Yu N, *et al*., Handwritten digits recognition base on improved LeNet5[C]. *Chinese Control and Decision Conference*, (2015), pp. 4871-4875.
[19]Yann LeCun, Learning Invariant Feature Hierarchies, in Fusiello, Andrea and Murino, Vittorio and Cucchiara, Rita (Eds), *European Conference on Computer Vision (ECCV 2012)*, 7583 (2012), pp.496-505
[20] T. Wang, *et al*., End-to-end text recognition with convolutional neural networks, *Proceedings of the 21st International Conference on Pattern Recognition(ICPR 2012. Tsukuba Science City, JAPAN)*, (2012)
[21]Shi B, *et al*., An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39 (2015), 11, pp. 2298-2304.
[22]He K, *et al*., Deep Residual Learning for Image Recognition[C]. // *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). IEEE Computer Society*, (2016).
[23]Ioffe S, *et al*., Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. *ICML 2015*, (2015).
[24]Hochreiter S, *et al*., Long short-term memory.*Neural Computation*, 9 (1997), 8, pp. 1735–1780.
[25]Graves A, Gomez F. Connectionist temporal classification:labelling unsegmented sequence data with recurrent neural networks[C]// *International Conference on Machine Learning*. 2006.
[26] Kingma D，Ba J.Adam：A method for stochastic optimization[J], *Computer Science* (2014).