9th International Conference on Information Science and Technology (ICIST), Hulunbuir, Inner Mongolia, China; August 2-5, 2019

Research on Offline Handwritten Chinese Character Recognition Based on Deep Learning

Qiuyun Hao, Xiaoming Wu, Sen Zhang, Peng Zhang, Xiaofeng Ma, Jingsai Jiang
Qilu University of Technology (Shandong Academy of Sciences), Shandong Computer Science Center (National Supercomputer Center in Jinan), Shandong Provincial Key Laboratory of Computer Networks
Jinan, China
haoqy@sdas.org

***Abstract*—Offline Handwritten Chinese Character Recognition (HCCR) is not only a research difficulty, but also a research hotspot. With the rapid development of deep learning, offline HCCR which based on deep learning has been widely studied and made a Breakthrough Development in methods and performance. CASIA-HWDB1.1 handwritten Chinese character database which was collected by Institute of Automation, Chinese Academy of Sciences is adopted in the paper. In the paper, tensorflow software platform is used, and the structure of convolution neural network (CNN) is optimized. The neural network model used in the paper has been trained and tested. It can be concluded that the recognition rate of offline HCCR has been improved. The training model has been successfully applied to the off-line HCCR system based on Android, and good results have been achieved.**

Keywords—Deep Learning; Offline HCCR; CNN; Tensorflow

I. Introduction

Among many languages, the Chinese character is the language with the largest number of users, the most concise and rich information meaning. Therefore, Chinese character recognition has very important research value, and has become a research hotspot. At present, the research of Chinese character recognition technology is mainly divided into printed Chinese character recognition (PCCR) and HCCR. However, handwritten Chinese characters have many characteristics, such as many categories of Chinese characters, very complex font structure, and a lot of changes in font shape, many similar characters, writing arbitrariness and strong differences in writing style. In addition, different people have different writing styles, and the same person's handwritten Chinese characters are very different in different writing environments and writing methods. Therefore, due to the characteristics of HCCR, it has always been a research difficulty in the field of pattern recognition. HCCR includes offline HCCR and online HCCR. In online HCCR, the handwriting process of Chinese characters is recorded by an intelligent device which is connected to a computer. It is based on strokes. According to the unique differences of Chinese characters, such as different stroke combinations, different relative position relations of strokes and different writing order of each Chinese character, the online Chinese character is recognized[1, 2]. For offline HCCR, the computer equipment scans and reads the text images of Chinese characters that have been written on the media. The Chinese characters in the image can be recognized automatically. Unlike online HCCR, offline handwritten Chinese characters do not have the timing information of writing trajectory in the process of writing. It means that Chinese characters are recognized from two dimensional images with missing stroke order information. Because it is based on the data information of the image, the available effective feature information is relatively small, which greatly increases the difficulty of offline HCCR. In addition, due to the characteristics of handwritten Chinese characters, offline HCCR has become a challenging problem.

This work was supported by the National Natural Science Foundation of China (Grant No. 61601267), by the S&T Innovation Program of Shandong Academy of Sciences, and by the International S&T Cooperation Program of Shandong Academy of Sciences (2019GHPY18).

Generally speaking, the traditional HCCR methods are mainly based on preprocessing, feature extraction [3, 4] and classifier. HCCR system can be classified according to different feature extraction methods and classifiers. According to the feature extraction method, it can be divided into two categories: the statistical feature based on Chinese character font and the structural feature based on Chinese character font. It can also be used in combination. For offline HCCR, Gabor feature [3] and gradient feature [4] are two better direction feature extraction methods at present. The most commonly used classifier models include modified quadratic discrimination function (MQDF) [5], support vector machine (SVM) [6], hidden markov model (HMM) [7], discriminative learning quadratic discrimination Function (DLQDF) [8], learning vector quantization (LVQ), etc. After years of development, great progress has been made in HCCR. With the rise of deep learning, the traditional recognition algorithms are constantly subverted by deep learning theory [10-12]. Especially, there are many breakthroughs in deep learning models such as CNN [13-14], deep belief network (DBN) [10], Stacked auto-encoder (SAE) [15], deep recursive neural network (DRNN) [16]. In the fields of image recognition and computer vision, a large number of breakthroughs have been made in these deep learning models. These related methods have been gradually applied to the field of handwritten character recognition, and many breakthroughs have been made.

II. Principle of CNN

In 1987, Rumelhart et al. completely proposed the back propagation (BP) algorithm [17], which systematically solved the learning problem of the hidden units' connection weight in multi-layer networks. And it gave a complete derivation in mathematical books. This was a milestone in the history of neural networks. BP arithmetic became popular rapidly, and the second climax of neural network was set off. Then, in

1989, LeCun et al. used it in multi-layer neural networks to recognize handwritten numbers [18]. Until 1998, LeNet-5 model [14] was put forward by LeCun et al., which marked the formal formation of CNN. Lecun et al. designed and trained the earliest convolutional neural network by gradient descent method for the first time. In the test of some pattern recognition tasks, the best recognition rate was obtained. Until now, the pattern recognition system which based on CNN has been able to achieve the best performance. Especially, Lecun's pattern recognition system for handwritten character recognition, as a machine learning platform server, is for people to study and use. Under GPU accelerated hardware conditions and application background of large data recognition, deep learning and CNN have developed rapidly.

CNN is a multi-layer neural network, which is good at dealing with machine learning problems related to images, especially large images. CNN can continuously reduce the dimension of image recognition problem with huge data volume by a series of methods, and finally make it able to be trained. Convolution is used by CNN to simulate feature differentiation. Moreover, it reduces the order of network parameters by sharing and pooling the weights of convolutions. Finally, the traditional neural network is used by CNN to accomplish classification tasks. Usually, the following layers are included by CNN.

- Input layer. In the layer, the original image data is preprocessed, including removing mean, normalization and principal component analysis (PCA)/whitening.
- Convolutional layer. Each convolution layer of CNN consists of several convolution units. The parameters of each convolution unit are optimized by BP algorithm. The purpose of convolution is to extract different features of the input data. Some low-level features may only be extracted by the first convolution layer, such as edges, lines and angles. More complex features are extracted iteratively from the lower features by more layers of networks.
- Rectified Linear Units (ReLU) layer. ReLU are used by the layer's activation function. The excitation function of CNN is generally ReLU. The characteristics of ReLU are fast convergence, simple gradient calculation, but relatively fragile.
- Pooling layer. Generally, many features with large dimension are obtained after convolution layer. These features are cut into several regions by the pooling layer, and the maximum values or the average values are taken. New features with smaller dimensions are obtained. Briefly, the pooling layer is mainly used to compress data to reduce over-fitting. If the input data is images, the main function of the pooling layer is to compress images.
- Fully Connected layer. In the layer, all local features are combined into global features to calculate the final score of each category.

LeNet-5 model is the most typical CNN model, as shown in Fig. 1. The model consists of convolution layer, pooling layer

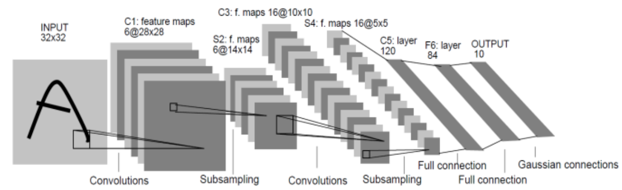

Fig. 1. The structural diagram of LeNet-5 convolution model[20]

and fully-connected layer. Among them, the convolution layer and the pooling layer are composed of several convolution groups. The features are extracted layer by layer. Finally, classification is completed through several fully-connected layers. Convolution is used by CNN to simulate feature differentiation. Generally speaking, convolution is used by CNN to simulate feature differentiation. It reduces the order of network parameters by sharing and pooling the weights of convolutions. The traditional neural network is used by CNN to accomplish classification tasks.

III. DATABASES OF HANDWRITTEN CHINESE CHARACTER(HCK2000，CASIS-HWDB1.1)

With continuous development of HCCR, some researchers began to establish some large-scale standard databases. In order to facilitate the research of handwritten Chinese character recognition, the establishment of handwritten Chinese character database has been paid more and more attention by domestic scholars. At present, there are several authoritative handwritten Chinese character databases in China, including HCL2000 handwritten Chinese character database by Beijing University of Posts and Telecommunications, CASIA series handwritten Chinese character database by Institute of Automation, Chinese Academy of Sciences [19], HIT-MW handwritten Chinese character database and HIT-OR3C handwritten database collected by Harbin Institute of Technology, SCUT-COUCH handwritten Chinese character database collected by South China University of Technology [20]. These databases provide a basis for the research of HCCR.

A. HCL2000 Database

HCL2000 is the largest offline handwritten Chinese character database in China. HCL2000 database is one of the key research projects funded by 863 National Program. It wasorganized and established by Laboratory of Pattern Recognition and Intelligent System, Beijing University of Posts and Telecommunications. HCL2000 is written by 1000 people, with 1000 sets of samples, each containing 3755 commonly used Chinese characters from the GB2312-80 first class font library. In order to study the related factors, the database not only records the samples of handwritten Chinese characters, but also records age, occupation, educational level, background and other information of 1000 people. At present, the database is free and open to the outside world.

B. CASIA-HWDB1.1 Database

In order to support the research of HCCR, CSAIA-HWDB database has been constructed successively by Institute of Automation, Chinese Academy of Sciences. In 2010, three offline handwritten Chinese character databases were published, including CASIA-HWDB1.0, CASIA-HWDB1.1 and CASIA-HWDB1.2. CASIA-HWDB 1.0 database contains 171 English numeric symbols and 420 sets of handwritten Chinese characters, each of which contains 3866 commonly used Chinese characters. CASIA-HWDB1.1 database and CASIA-HWDB1.2 database include 300 sets of Chinese characters, including 6763 Chinese characters in GB2312.

CASIA-HWDB1.1 database is selected as the database in the paper. The database consists of handwritten Chinese character images. It was written by 300 people, and all the sample images which were written by each person were saved in a file. The database includes 171 English numeric symbols and 3755 first-level Chinese characters of GB2312-80 character set. Each sample image is an 8 bit gray image, as shown in Fig. 2. In the sample library of each Chinese character, 240 samples were used for training and 60 samples were used for testing.

Fig. 2. The Sample example of CASIA-HWDB1.1

IV. Training and Testing of CNN Model

Tensorflow is selected as the development tool of deep learning in the paper. Tensorflow is a visual development tool for machine learning released by Google. It supports multi-CPU and multi-GPU parallel simulation, and also supports deep learning models such as CNN and RNN.

The CNN model which is adopted in the paper consists of input layer, four convolutional layers, four pooling layers and two fully connected layers, as shown in Fig. 3, Fig. 4 and Fig. 5. The input data of the input layer is 64*64 handwritten Chinese character images. The first convolution layer C1, as shown in Fig. 3, consists of 64 feature maps, each of the feature maps consists of 3*3 convolution kernels. There are (3*3+1)*64=640 training parameters and the size of each feature surface is 64*64. After the convolution layer C1, the pooling layer S2 performs the maximum pooling operation on the input feature map. The layer consists of 64 feature maps with size of 32*32. Each node in the feature map is connected with the corresponding feature map in the C1 layer with a sampling core of 2*2. There are a total of 1*64=64 offsets. The second convolution layer C3, as shown in Fig. 4, contains 128 feature maps. Each feature map consists of 3*3 convolution kernels, and the size of each feature surface is 32*32. The remaining convolutional layers and pooling layers are not described in detail. Among them, S4 contains 128 feature maps with size of 16*16, C5 contains 256 feature maps with size of 16*16, S6 contains 256 feature maps with size of 8*8, C7 contains 512 feature maps with of 8*8 and S8 contains 512 feature maps with size of 4*4. As shown in Fig. 5, after the pooling layer S8, the fully connected layer F9 straightens the matrix into 4*4*512 vectors. The number of output nodes in the layer is 8192. Then it passes through the fully connected layer F10, and the number of output nodes is 1024. Finally, it is sent to the output layer for classification, the classification number is 3755. Classifiers have many choices. The softmax regression classifier is adopted in the paper. Under tensorflow development environment, the designed CNN model in the paper is trained and tested. It can be concluded that the recognition rate of top1 can reach about 90%, and that of top3 can reach about 96%. In order to further analyze the results, the accuracy and loss value of the CNN model are obtained by using Tensor Board tool. As shown in Fig. 6 and Fig. 7, along

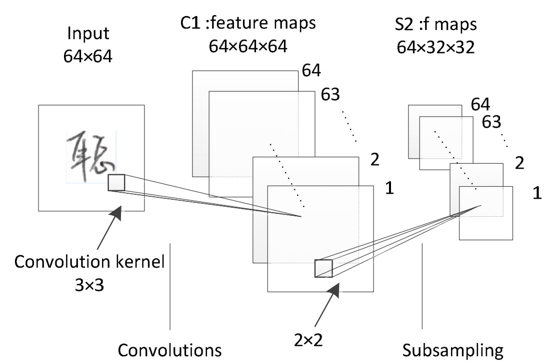

Fig. 3. Input layer, convolutional layer C1 and pooling layer S2

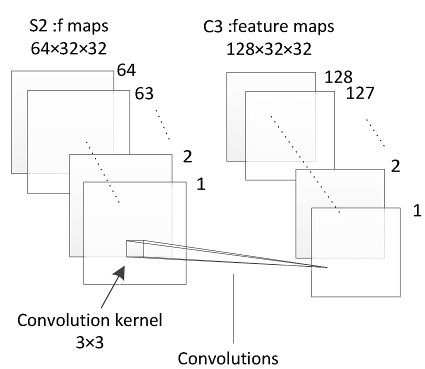

Fig. 4. Pooling layer S2 and convolutional layer C3

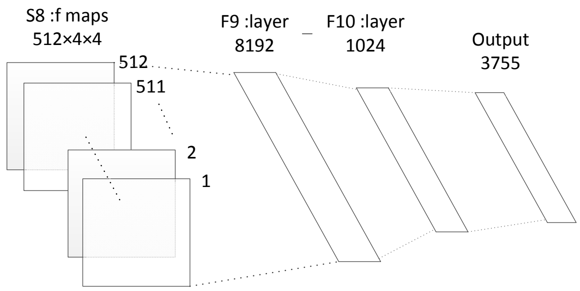

Fig. 5. Pooling layer S8, fully connected layer F9, F10 and output layer

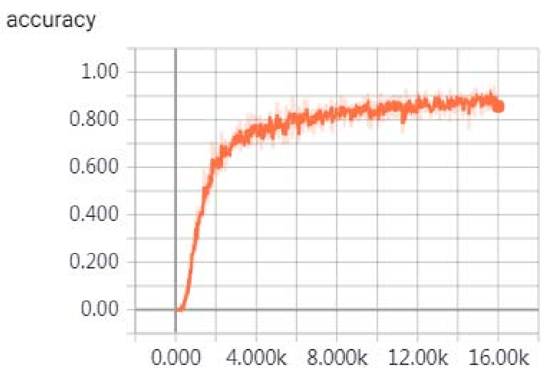

Fig. 6. The curve of iteration number and accuracy

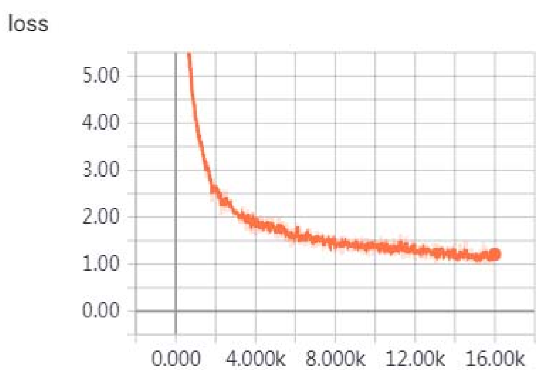

Fig. 7. The curve of iteration number and loss value

with increasing of the iteration number, the accuracy is gradually increasing and the loss value is gradually decreasing. The accuracy rose fastest between 0 and 2000 steps. The reason is that it is the initial stage. At the stage, the parameters of the model are constantly updated, the accuracy increases rapidly, and the loss value decreases rapidly. With the increasing of iteration number, the accuracy increases slowly, and gradually tends to be stable. But after 15000 steps, the accuracy of top 1 is close to 90%.

Based on the CNN model and CASIA-HWDB1.1 database, the recognition rate of the designed CNN model reaches 90%. The next step is the application of the trained model. In the paper, the trained model is solidified into PB file, and then the model is compressed. Finally, the model was successfully transplanted to an off-line HCCR system based on Android.

V. CONCLUSION

In order to meet the requirements of off-line HCCR system, CNN and CASIA-HWDB1.1 database are adopted in the paper. Through the analysis and structure optimization of CNN, the CNN model suitable for offline HCCR system is designed. In the paper, CASIA-HWDB1.1 handwritten Chinese character database is tested. It can be concluded that the accuracy of top1 is about 90%, and that of top3 is about 96%. The recognition rate of offline HCCR is improved. After the trained model by tensorflow is solidified into PB file, the corresponding model is compressed and successfully applied to the offline HCCR system based on Android.

REFERENCES

[1] Du J, Zhai J F, Hu J S, “Writer adaptation via deeply learned features for online Chinese handwriting recognition,” International Journal on Document Analysis & Recognition, 2017, pp.1-10.

[2] Zhang X Y, Bengio Y, Liu C L, “Online and Offline Handwritten Chinese Character Recognition: A Comprehensive Study and New Benchmark,” Pattern Recognition, • 2016, pp.348-360. •

[3] Ge Y, Huo Q, Feng Z D, “Offline recognition of handwritten Chinese characters using Gabor features, CDHMM modeling and MCE training,” In: Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, FL, USA: IEEE, 2002, pp.1053-1056.

[4] Liu C L, “Normalization-cooperated gradient feature extraction for handwritten charSacter recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, vol.29, no.8, pp.1465−1469.

[5] Kimura F, Takashina K, Tsuruoka S, Miyake Y, “Modified quadratic discriminant functions and the application to Chinese character recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1987, PAMI-9(1):149-153.

[6] Mangasarian O L, Musicant D R, “Data discrimination vianonlinear generalized support vector machines,” Complementarity: Applications, Algorithms and Extensions. US:Springer, 2001, pp.233-251.

[7] Kim H J, Kim K H, Kim S K, Lee J K, “On-line recognition of handwritten Chinese characters based on hidden Markov models,” Pattern Recognition, 1997, vol.30, no.9, pp.1489-1500.

[8] Liu C L, Sako H, Fujisawa H, “Discriminative learning quadratic discriminant function for handwriting recognition,” IEEE Transactions on Neural Networks, 2004, vol.15, no.2, pp. 430-444.

[9] Jin X B, Liu C L, Hou X W, “Regularized margin-based conditional log-likelihood loss for prototype learning,” Pattern Recognition, 2010, vol.43 , no.7, pp.2428-2438.

[10] Hinton G E, Salakhutdinov R R, “Reducing the dimensionality of data with neural networks,” Science, 2006, vol.313, no.5786, pp.504-507。

[11] Bengio Y, Courville A, Vincent P, “Representation learning: a review and new perspectives,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, vol.35,no.8, pp.1798-1828

[12] Schmidhuber J, “Deep learning in neural networks: an overview,” Neural Networks, 2015, vol.61, pp.85-117.

[13] LeCun Y, Boser B, Denker J S, Howard R E, Habbard W, Jackel L D, Henderson D, “Handwritten digit recognition with a back-propagation network,” In: Proceedings of Advances in Neural Information Processing Systems 2. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1990, pp.396-404.

[14] LeCun Y, Bottou L, Bengio Y, Haffner P, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, 1998, vol.86, no.11, pp.2278-2324.

[15] Ranzato M A, Poultney C, Chopra S, LeCun Y, “Efficient learning of sparse representations with an energy-based model,” In: Proceedings of the 2007 Advances in Neural Information Processing Systems. USA: MIT Press, 2007,pp.1137-1144.

[16] Hochreiter S, Schmidhuber J, “Long short-term memory,” Neural Computation, 1997, vol.9,no.8, pp.1735-1780.

[17] David E. Rumelhart, James L. McClelland, "Learning Internal Representations by Error Propagation," Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, 1987, pp.318-362.

[18] Lecun Y, Boser B, Denker J. S., Henderson D., et al, "Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation," 1989, pp.541-551.

[19] National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, "CASIA Online and Offline Chinese Handwriting Databases," http://www.nlpr.ia.ac.cn/databases/handwriting/Download.html.

[20] Hanyu Yan , Lianwen Jin , Viard-Gaudin, C. , Mouchere, H. , "SCUT-COUCH Textline_NU: An Unconstrained Online Handwritten Chinese Text Lines Dataset," 2010 International Conference on Frontiers in Handwriting Recognition (ICFHR), 2010, pp.581-586.