An Approach for Chinese Character Captcha Recognition Using CNN

Yang JIA, Wang FAN, Chen ZHAO, Jungang HAN
School of Computer Science and Technology, Xi’an University of Posts and Communications, Xi’an, Shaanxi 710061, China
Email: jiayang@xupt.edu.cn

Abstract. An approach using convolution neural network (CNN) to recognize Chinese Character Captcha captured from Tencent Security Center is proposed. 3500 commonly used Chinese characters are applied in the verification code system, so the problem is to detect, recognize and match the characters occurred in the two Captcha images. Firstly the characters in a Captcha image are analyzed and segmented, and each single character is a minimum unit in this paper. Secondly, a recognizing CNN is built and trained with a character library with 620,000 different character modes made by ourselves. Thirdly, the segmented characters from the Captcha image are input the network to finish recognition. The experiment result shows that the identification rate of the test data is as high as 99.4 percent after 25 iterations, and it is 1.75 percent higher when compared with another CNN model. It shows that the approach using CNN to do Chinese character Captcha recognition is feasible and it can provide a high level of accuracy.

1. Introduction

Human computer verification is a computer program that differentiates users from computers and people. It helps protect people from spam and malicious registration, and makes website more secure. Traditional man-machine validation method is to use graphics visual identification (Captcha) for verification, validation protocols have commonly used: image characters, such as letters and numbers shown in column (a)-(c) Figure 1. In China, Chinese characters have already been used as Captcha code as shown in column (d). There are only 26 letters and 10 numbers in the original library of a Captcha code system, while even the simple Chinese character Captcha library has 3500 commonly used characters. So recognizing a Chinese character Captcha is more difficult than letter and number Captcha[1]. Researchers try to make the Captcha strong enough to defend the attacks from the Internet, and people also think about approaches to break the Captcha. Breaking Captchas is an important way to find the defeats and it can help to improve the Captchas.
In most Captcha recognition studies, there are mainly two steps: feature extraction and classification\[2, 3\]. Anjali et al. [2] used zonal features of character to do character recognition, the experiment shown that the graph matching and the feature extraction were suitable for the typewritten fonts as the fonts are fixed and less in number. For the Captcha with a large amount of deformations the method is not applicable. Hussain et al. [3] firstly cropped the letters and numbers from the Captcha and trained a neural network on individual cropped characters, then extracted features were sent to ANN to identify the character one by one, the highest recognition rate is 97.25%. In recent years, deep learning technology is also used to do Captcha recognition. Lv et al. [4] proposed a Convolution Neural Network (CNN) based approach to improve the recognition accuracy of Chinese Character CAPTCHAs with deformation. Geetika [5] used CNN and recurrent neural networks instead of using the conventional approach of first cleaning, segmenting the image, and recognizing the individual characters. Experiment result indicates that a high accuracy could be got in letter and number Captcha with the method. Shen [6] proposed a multi-scale Corner based Structure Model to recognize Touclick Captchas.

As the improvement of hardware capability and algorithms, CNN shows strong capability in different pattern classification. In image processing field, with the usage of GPUs to speed up the training process, it is more convenient to train a large number of images. Caffe, Mxnet or TensorFlow framework are mostly used in Captcha recognition, while Keras framework [7] is rarely used. In this paper Keras neural network framework is adopted to built a suitable Captcha code recognition network to improve the accuracy. Keras is a simple and modular neural network library based on Theano and Tensorflow deep learning library. A new network structure based on Keras is designed to recognize Chinese character Captcha from the Tencent Security Center [8] in Keras network framework, as shown in column (e) Figure 1. The whole algorithm can be divided into four parts: character library building, character segmentation, network building and predicted probability matching. The flowchart of the algorithm is shown in Figure 2.

Details of the algorithm are described as follows.
2. Chinese Character Library Building
Font is the main factor causing the difference of Chinese characters. So 59 different fonts are selected from the Microsoft font library (10 fonts) and Foundertype font library (49 fonts). The Chinese character library consists of 3500 Chinese characters with different fonts, deformation[9], rotation, and noise. An individual character is 32\times32 pixels, and there are totally 620,000 characters in the character library. Some examples are shown in Figure 3.

In a row, parts of the deformations of one character are shown there, and four examples are shown in four rows.

3. Preprocessing of Chinese Character CAPTCHA
In this paper, 500 Chinese character Captchas are collected from Tencent Security Center[8]. As shown in Figure 4, column (a)-(c) are three groups of Captcha. Users need to follow the characters displayed in the first row to click four characters in the third row to pass the test. Based on the Captcha pattern, individual characters are segmented firstly.

The segmentation algorithm based on thresholding and morphological analysis is simple and fast. Following segmentation operations were performed:
(a) Image preprocessing: due to the lack of hue, the image is transformed into gray image and red channel of the RGB image is chosen as the gray image.
(b) Individual character region segmentation: in the captcha all the black characters have white border. While in some images, because of the whole white background the white border cannot be utilized for
segmentation. Just dark color of the characters is used to do segmentation. Firstly, thresholding is used to get the character. In this step, our aim is to get the character region and remove the background, so appropriate over segmentation is acceptable. The a strong dilatate operation is done to cover the individual character region. Then minimum bounding rectangles of the dilated foreground are treated as the character region.

(c) Character segmentation: Each minimum bounding rectangle is treated as an ROI and with less background interference the character is segmented with a thresholding method. The segmentation result is shown in row (2) and (4) in Figure 4.

![Figure 4](image)

Figure 4  Segmented Chinese characters from the Captcha. Three groups are shown in column (a)-(c). The first and the third rows are the Captcha images. The second and the fourth rows are the segmented characters.

The segmentation method is simple and fast, while it still needs to be explained that over segmentation (labeled red circle II, III, IV, V in Figure 4) and segmentation leakage (labeled red circle I, VI in Figure 4) occur occasionally. Because in some fonts, some strokes are thin and the color is not dark enough and those parts would be missed easily. If we try to keep the strokes, over segmentation will occur. So a balance need to be kept and these problems will be solved by the trained deep neural network model.

4. Neural Network Model Building

The character recognition network model is built with Keras[7]. The CNN network structure for deep learning is as shown in Figure 5. Input of the network is a three channel 32×32 character image. Each convolution layer is followed by an activation layer using ReLu. Input and output size of a convolution layer are set the same. Four pooling layers are designed to subsample the resolution of the feature maps in the previous layer with a 2×2 kernel. Then a Flatten layer is used to reduce each vector to one dimension. The flattened output is then fed into a fully connected (dense) layer. The final layer is fully connected to the previous dense layer with a softmax activation. Details can be found in Figure 5.

The network is used to train a Captcha recognition model. In order to improve the performance of recognizing, the CNN network structure is designed specially, and the network structure has significant advantages: (1) Each convolution layer uses smaller convolution kernels (3×3) to capture characters with complex stroke features. (2) More layers are used to capture the deep structure information of the characters in two images.

Considering the complexity of the Chinese characters, the charateres in the larger image in row (3) Figure 4 and the small image in row (1) Figure4 are predicted with the neural network separately, and then Euclidean distance of the two prediction matrix is calculated and checked for a match.
Figure 5  A CNN architecture for Chinese character Captcha recognition. This architecture consists of 10 convolution layers, 4 pooling layers and 1 fully-connected layer. Detailed descriptions are given in the text.

The softmax function is as Eq. (1):
\[
\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad (j = 1, 2, 3, ..., K)
\]  

K is the number of the character categories, and here K is 3500.

As shown in Figure 6, \( z \) is input of the fully connected layer, \( z = \sum_j w_j x_j + b \). \( x_j \) is input, \( w_j \) is weight, and \( b \) is bias. Each category has a \( z \), and the softmax function is used to calculate the probability of the category. For example, a character will have 36 categories after the calculation of the network and each category has its own \( \sigma(z) \). And \( \sigma(z) \) is set the maximum of the 36 \( \sigma(z) \). The loss function is as Eq. (2) shows:
\[
C = -\sum x \ln(a) + (1 - y) \ln(1 - a)
\]  
\( a = \sigma(z) \) is true label, \( y \) is the target. Derivative of \( C \) is:
Then the weights are optimized.

\[
\frac{\partial c}{\partial w_j} = \frac{1}{n} \sum_k X_j (\sigma(z) - y)
\]

(3)

\[
\frac{\partial c}{\partial b} = \frac{1}{n} \sum_k (\sigma(z) - y)
\]

(4)

5. Experiment

We perform experiments on the Chinese Character Captcha library which consists of 620,000 images of characters as shown in Figure 3. The designed network is run in Spyder 3.1.2 on a server with an Intel Core i7-5820k (R) central processing unit (3.30 GHz), 32 GB memory and a GeForce GTX 970 GPU with 4 GB graphics memory. The segmented individual character picture is normalized to 32×32 pixels to be the input of the network.

| Table 1 Training result | Table 2 Comparison of the Training result | Table 3 Click recognition accuracy |
|-------------------------|------------------------------------------|----------------------------------|
| Number of iterations   | Accuracy | Methods | Accuracy | Methods | Accuracy | Operations | Results |
| 5  | 81.2%    | Methods in [4] | 97.72%   | Methods in this paper | 99.47% | Segmented images | 500 images |
| 10 | 93.24%   |       |          |          |        | Recognition Accuracy | 87% |
| 20 | 97.71%   |       |          |          |        | Recognized images | 437 images |
| 25 | 99.47%   |       |          |          |        | Recognition Accuracy | 92.8% |

In the network Adam optimizer proposed in [10] is used. Memory consumption of the optimization methods is not high, and it is very suitable for problems with large data and parameters. The hyper-parameter setting is set with the default value. The learning rate lr is set to 0.0001, the remainder uses Adam’s default value, the selected block size (batch size) is 384. GPU is used to speed up the training process, and the maximum number of training iterations is 25. Training results are in Table 1.

The training result is compared with the results obtained with the method in [4], and the experimental results are shown in Table 2.

Simulation of the Captcha click experiment is tested and the result is shown in Table 3.

6. Conclusion

In this paper, a CNN framework based on Keras is designed to effectively improve the recognition rate of Chinese character Captcha. A Chinese character library consists of 620,000 character images (3500 characters) with different fonts, distortion, rotation, and noise are built. Strong robustness in prediction can be guaranteed with the library. The accuracy of the network test experiment is almost 100%, while the recognition accuracy of the click test still needs to be improved. In the future, a larger Chinese character library with some samples of the segmented characters will be built to improve the performance of character image recognition. On the other hand, mixed character Captcha will be studied and the further goal is to distinguish a Captcha mixed with letters, numbers and Chinese characters.

References

[1] YU J, MA X, HAN T. Usability Investigation on the Localization of Text CAPTCHAs: Take Chinese Characters as a Case Study [J]. 2016,

[2] CHANDAVALE A, DEDGAONKAR S, SAPKAL A. An Approach for Character Recognition Using Pattern Matching with ANN [J]. International Journal of Scientific & Engineering Research, 2012,

[3] HUSSAIN R, GAO H, SHAIKH R A, et al. Recognition based segmentation of connected characters in text based CAPTCHAs; proceedings of the IEEE International Conference on Communication Software and Networks, F, 2016 [C].

[4] LV Y, CAI F, LIN D, et al. Chinese character CAPTCHA recognition based on convolution neural network; proceedings of the Evolutionary Computation, F, 2016 [C].

[5] GARG G. Neural Network CAPTCHA Crackers [D]; San Jose State University, 2015.
[6] SHEN Y, JI R, CAO D, et al. Hacking Chinese Touclick CAPTCHA by Multi-Scale Corner Structure Model with Fast Pattern Matching [M]. Proceedings of the 22nd ACM international conference on Multimedia. Orlando, Florida, USA; ACM. 2014: 853-6.

[7] WILLIAMS M L. Getting Started with Deep Learning using Keras and Python [J]. 2016,

[8] Tencent Security Center. https://aq.qq.com/v2/uv_aq/html/reset_pwd/pc_reset_pwd_input_account.html?v=4.0&old_ver_account=376136849

[9] WANG T, B EGH J. Multi-layer CAPTCHA Based on Chinese Character Deformation; proceedings of the International Conference on Trustworthy Computing and Services, F, 2013 [C].

[10] KINGMA D, BA J. Adam: A Method for Stochastic Optimization [J]. Computer Science, 2014,