CNN for breaking text-based CAPTCHA with noise

Kaixuan Liu, Rong Zhang, Ke Qing
Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei, China
Key Laboratory of Electromagnetic Space Information,
Chinese Academy of Sciences, Hefei, China

ABSTRACT

A **CAPTCHA** ("**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**uman **A**part") system is a program that most humans can pass but current computer programs could hardly pass. As the most common type of CAPTCHAs , text-based CAPTCHA has been widely used in different websites to defense network bots. In order to breaking text-based CAPTCHA, in this paper, two trained CNN models are connected for the segmentation and classification of CAPTCHA images. Then base on these two models, we apply sliding window segmentation and voting classification methods realize an end-to-end CAPTCHA breaking system with high success rate. The experiment results show that our method is robust and effective in breaking text-based CAPTCHA with noise.

Keywords: CNN, sliding window, synthetic data, CAPTCHA

1. INTRODUCTION

As the most common CAPTCHA, text-based CAPTCHA [1] has been widely used in a large number of sites. Text-based CAPTCHA is a kind of visual Turing test. A text-based CAPTCHA usually shows randomly combination of characters that include 26 upper and lower case letters (A..Z, a..z), 10 digits (0..9). The use of wide variance of shapes, transformations and distortions or even add lines and background noise [2-3] to write letters make the recognition of text-based CAPTCHA a complex problem for computer while humans are easily able to recognize.

Traditionally, defeating a text-based CAPTCHA test requires two states: segmentation and recognition. Research [5] shows that the segmentation step is more hardly than classification step when solving text-based CAPTCHAs, because text-based CAPTCHAs were usually designed to be segmentation-resistant. Typically, some text-based CAPTCHA image like to be added with noise that segmentation algorithm can not splits the image into segments contain individual characters. It is hard to design an effective way to remove various line noise in a CAPTCHA, some traditional ways like vertical histogram segmentation [4] and dilate & erode algorithm [5] have been used in segmenting CAPTCHA image with noise, while these algorithms are too simple to adapt the variation of noise. This will result in a low SSR (success segmentation rate) and the final accuracy of CAPTCHA test can't reach a high level. On the other hand, CNN, a kind of deep natural network, has reached state-of-the-art performance in image classification, segmentation and object detection et al [6-9]. CNN has powerful ability in feature learning with large training dataset. It provides us a new way in breaking text-based CAPTCHA with complex noise.

In consideration of these conditions proposed above, in this paper, we design an effective end-to-end text-based CAPTCHAs attacking framework which connects tow CNN models together. Particularly, several important methods are applied in this framework and keep a high performance in breaking text-based CAPTCHA with noise. First, sliding window method makes it easy for CNN to find the locations of complete single characters. Then we design synthetic algorithm to enlarge training data and this make trained model more robust with the variance of test data. At last, when classifying single character images, we enlarge one image into 30 images by using the synthetic algorithm and then apply voting principle to choose the one with highest accurate times as final result. This method against except errors when classify one character just one time. The frame were tested on souhu CAPTCHA (a type of CAPTCHA with line noise, showed in Figure 1.) and reaches 96.82% success segmentation rate (SSR), 86.69% single characters recognition rate (SRR) and 67.71% CAPTCHA test success rate.

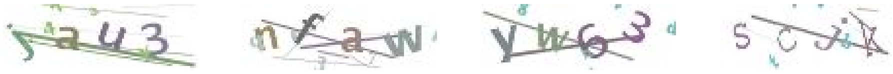
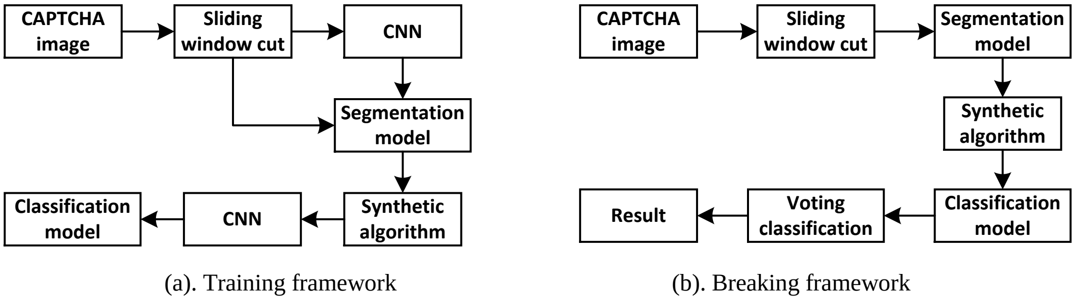


Figure 1. Examples of sohu CAPTCHAs

Ninth International Conference on Digital Image Processing (ICDIP 2017), edited by Charles M. Falco, Xudong Jiang
Proc. of SPIE Vol. 10420, 104202V · © 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2281743

2. END-TO-END CAPTCHA BREAKING SYSTEM

In this section, we will make a detail description about our CAPTCHA breaking system. When accomplishing the CAPTCHA breaking system, our work can be divided into two parts - training models and establish an end-to-end breaking framework. As shown in Figure 2, we first trained a segmentation model, then the model was used to cut labeled CAPTCHA images into single character images, these images were used for the training of classification model. Synthetic algorithm was applied for enlarging training data. When testing a CAPTCHA image, we first cut the image into segments with single characters, then classification model can give the result of every character. Synthetic and voting method help the breaking framework be more robust to the variance of characters. The sliding window cut and synthetic algorithm which are applied in both training and breaking process are presented in 2.1 and 2.2. In 2.3, we introduce our models training method. In 2.4, an essential algorithm for recognize single character, voting classification, is described.



(a). Training framework

(b). Breaking framework

Figure 2. CAPTCHA breaking system. (a) Training framework. (b) CAPTCHA breaking framework.

2.1 SLIDING WINDOW CUT

As the basic method in segmentation, sliding window cut method provides an effective framework to segment CAPTCHA images into individual characters. The window we set has same high with CAPTCHA and the width is 25 pixels. We slide window from left to right with the steps of 3 pixels to extract patch of images which were latterly used for the training of segmentation and classification.

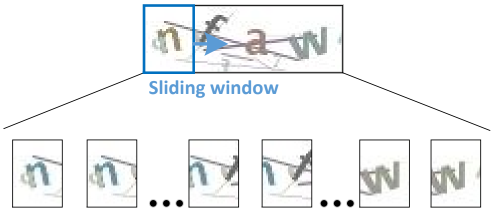


Figure 3. Sliding window cut

2.2 SYNTHETIC ALGORITHM

Because we can’t get labeled sohu CAPTCHA images straightly from websites, and it is very time-consuming that any data for training must be labeled with human. Taking the lacking training data problem into consideration, we have designed synthetic algorithm that can simulate the transformation of sohu CAPTCHA images to enlarge training data for classification. Our synthetic algorithm can be divided into three steps as follows:

1. *Resize* – we resize the image with liner interpolation when get the input image with single character. The size of new image is random form 0.7 to 1.3 times of original image.

2. *Rotate* – The range of rotation we set is from -15 to 15 degree, we randomly choose an angle in this range and rotate the image.

3. *Warp* – warp is used for making some distortion with image. After rotation, the image will be mapped from rectangle to a quadrilateral which is randomly set.

Figure 4 shows an example of transforming characters with synthetic algorithm. One character image can be transformed into many images and these images can be applied in the classification model and make the model more robust to the variance of CAPTCHA characters.

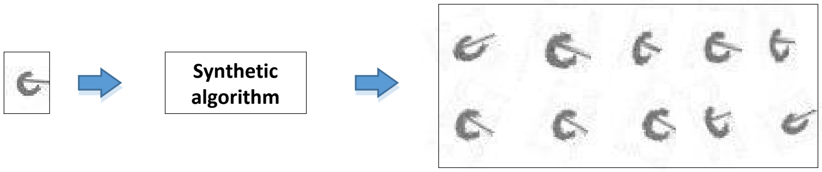


Figure 4. Example of synthetic algorithm

2.3 MODELS TRAINING

The training framework is showed in (a) of Figure 2, the sliding window cut method were used to segment a CAPTCHA image into mini patches and label them according to which image contain a complete character. The parameters of CNNs for training are showed in Tabel 1. The labeled patches were used for training segmentation model which can tell where the input image contains single complete character. After of this, we make use of labeled CAPTCHA images and train segmentation model to accomplish the task of training classification model. First, we find the images with complete single characters after segmentation. Then the single characters can be labeled according to the label of character image. When have acquired the labeled single characters, we apply synthetic algorithm to enlarge training data, this method can improve the performance of classification model obviously.

Table 1. Parameters of CNN layers. CNN-seg is for the training of segmetation, CNN-classi is for the training of classification. ks – kernel size, ss – stride step, softmax is used in the last layer.

Layer	**Settings**
Conv-1	ks=5, ss=1, nMap=20
Pool-1	ks=2, ss=2
Conv-2	ks=5, ss=1, nMap=50
Pool-2	ks=2, ss=2
Full-1	nNode=500
Full-2	CNN-seg nNode=2
	CNN-classi nNode=30

2.4 VOTING CLASSIFICATION

When the test CAPTCHA image is segmented into single characters, the single characters were recognized one by one according to their locations. For example, when the first character is to be recognized, we apply synthetic algorithm to expand number N times then put images into classification model. There are 30 output nodes of classification model, which stand for 30 characters which showed in sohu CAPTCHA. We make count of the results of every input image:

$$m_i = \sum_{j=1}^{N} I(f(x_j) = i) \tag{1}$$

m_i stands for the times of recognized as i , $f(x_j)$ is the result of j_{th} image. if the result is i ,number m_i plus 1, so function I is:

$$I(f(X),Y)=\begin{cases}1, & f(X)=Y\\ 0, & f(X)\neq Y\end{cases} \tag{2}$$

At last, we calculate the max m and the corresponding character is the final result.

3. EXPERIMENT & RESULTS

In this section, we firstly introduce the dataset for training and test. After that, we have established an end-to-end CAPTCHA breaking framework based on CNN models. Then on the basis of breaking framework, a set of experiments have been designed to evaluate our method.

Dataset: In the process of training segmentation model, we get 10,000 labeled data that include 8,000 images with complete characters and 2,000 images without complete character. The test data for segmentation is 400 sohu CAPTCHA images which got from the website. When training the classification model, we have 1200 labeled sohu CAPTCHA images then 4800 character images acquired after segmentation and these images are original training data for classification model. Furthermore, the synthetic algorithm has been used to enlarge training dataset, about 140,000 images were used for the training of classification model at last.

Experiment Results: As shown in Figure 2 (b), our end-to-end CAPTCHA framework make segmentation first and then images with single characters were classified by classification model. Finally, we get the CAPTCHA's results. The framework segments CAPTCHA images with 96.82% success rate. When classifying character images, the synthetic algorithm plays important role in improving model training and voting classification performance. In order to prove that, we have designed several experiments and test on the same dataset (400 sohu CAPTCHA images). Performance of our experiments are showed in Table 2.

Table 2. Performance of our different experiments. Data were tested on the computer with Core i7 processor of 3.4GHz, 16.0 GB RAM, Linux 16.04 x 64. 1x Training Data is 4800 original character images were used to train classification model. 10x or 30x training data stands for the training data has been enlarged up to 10 or 30 times of original data. 30x VC is that an image is transformed into 30 images for voting classification.

Experiments	Character accuracy	CAPTCHA success
1x Training Data	49.36%	10.48%
10x Training Data	65.3%	22.4%
10x Training Data + 30x VC	85.55%	65.44%
30x Training Data + 30x VC	**86.69%**	**67.71%**

Taking the experiment results into consideration, we find that, synthetic algorithm makes sufficient improvement for the success recognition rate of sohu CAPTCHA. The final CAPTCHA success rate proved that our CAPTCHA breaking system is effective in breaking text-based CAPTCHA and robust with the line noise and characters' transformation.

4. CONCLUSION

In this paper a novel approach to break text-based CAPTCHA has been presented, which the sliding window cut is an effective way to segment CAPTCHA images with CNN, then synthetic algorithm and voting classification method realize a robust classification model. Our end-to-end CAPTCHA breaking system provides a new frame work in attacking text-based CAPTCHA based on CNN. As future work, we will make further study on synthetic algorithm to solve the issue of lacking training data when attacking various text-based CAPTCHAs using deep learning. Further more, we will keep on studying more effective algorithm to segment adhesive text-based CAPTCHA.

REFERENCES

[1] Simard, P. Y. "Using machine learning to break visual human interaction proofs (hips." Advances in neural information processing systems 17: 265-272 (2005).

[2] El Ahmad, Ahmad Salah, Jeff Yan, and Lindsay Marshall. "The robustness of a new CAPTCHA." Proceedings of the Third European Workshop on System Security. ACM, (2010).

[3] Conti, Mauro, Claudio Guarisco, and Riccardo Spolaor. "CAPTCHaStar! A novel CAPTCHA based on interactive shape discovery." International Conference on Applied Cryptography and Network Security. Springer International Publishing, (2016).

[4] Yan, Jeff, and Ahmad Salah El Ahmad. "A Low-cost Attack on a Microsoft CAPTCHA." Proceedings of the 15th ACM conference on Computer and communications security. ACM, (2008).

[5] Bursztein, Elie, Matthieu Martin, and John Mitchell. "Text-based CAPTCHA strengths and weaknesses." Proceedings of the 18th ACM conference on Computer and communications security. ACM, (2011).

[6] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. (2012).

[7] Chen, Liang-Chieh, et al. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." arXiv preprint arXiv:1606.00915 (2016).

[8] LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." Neural computation 1.4 (1989): 541-551.

[9] Wang, Tao, et al. "End-to-end text recognition with convolutional neural networks." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, (2012).