Decoding Captcha using Machine Learning for Identifying Criminal Attempts

Dr. T. Venkat Narayana Rao
Professor, CSE, SNIST
Sreenidhi Institute of Science and Technology
Hyderabad, India

A. Sai laksmi
Student , CSE, SNIST
Sreenidhi Institute of Science and Technology
Hyderabad, India

Abstract : Completely Automated Public Turing Tests to Tell Computers and Humans Apart (CAPTCHAs) is a computer program which is meant to set humans and machines apart. In simple terms it is a system which identifies whether the user is a human or a machine. We use Captcha very often in our daily basis when we try to access internet. Captcha is meant for security. Decoding or recognizing the patterns of Captcha can help us know about any morally offensive activities taking place prior. Also we can stalk the systems of those who tend to misuse their powers and cause harm to the country. They are designed in such a way that they are solved by humans and unsolvable by machines.We can use convolutional neural networks and recurrent neural networks instead of the conventional methods of CAPTCHA breaking based on segmenting and recognizing a CAPTCHA. This paper mainly focuses on training a machine to decode Captcha which can be used to identify the criminal activities stalking the criminal activities in advance as well.

***Keywords**: Captcha, Convolutional Neural Networks(CNN)*

*****

1. Introduction

Captcha (Completely Automated Public Turing test to tell Computers and Humans Apart) was invented in 1997[1]. This was in the form of letters which are present in a distorted image which sometimes also had digits in it. The term was coined in 2003 by Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford[2]. The basic use of Captcha is to make sure that user involved in online transactions is a human being but not a machine. We often find the letters in a Captcha are blurry. This is because the letters or the words are retrieved from the scanned texts and are mistranslated by Optical Character Recognition software, so that a machine can't understand it. We find a text box below the Captcha which is meant to rewrite the letters present in the Captcha. In 2014, a Google analysis found that artificial intelligence could crack even the most complex CAPTCHA and reCAPTCHA images with very high accuracy of about 99.8%. It takes the average person approximately 10 seconds to solve a typical Captcha[3]. Security is the primary advantage of using Captcha. It does provide security by distinguishing human from machine so that we can make sure that the online transactions made are not fake. The human and machine can be set apart by using Captcha because a machine cannot recognize or identify the sequences or letters in Captcha as a human does. Unless and until a Captcha is filled in the textbox, the user cannot proceed further in accomplishing the task. So this way, Captcha blocks the machine from accessing the task. In this way Captcha provides security.

When Captchas came into existence, there was no methods which were capable of solving them. But with evolution of technology, there came into existence the neural networks to be more precise convolution neural networks which can be used to decode the Captcha for identifying intruders. The reason behind using CNN is to gain more accuracy. A popular deployment of CAPTCHA technology, reCAPTCHA, was acquired by Google in 2009[4].

These newest iterations have been much more successful at warding off automated tasks[5].

When it comes to Machine Learning, Artificial Neural Networks perform really well. Artificial Neural Networks are used in various classification tasks like image, audio, words. Different types of Neural Networks are used for multiple different purposes, for example, to predict sequence of strings(words), we apply Recurrent Neural Networks, while for classifying an image, we apply Convolution Neural Network. In this, we are going to build basic building block for CNN.

Apply Convolution Neural Network (CNN)

CNNs use a variation of multilayer perceptrons designed to process minimal preprocessing i.e. figure 1. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics[6].

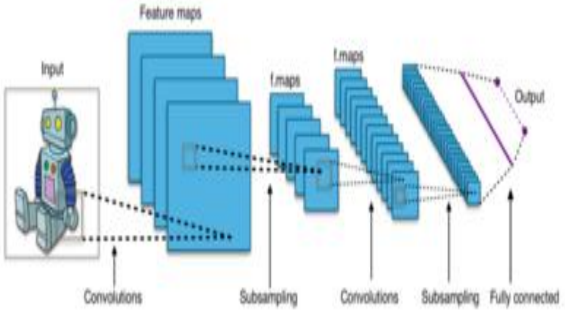


Figure 1. CNN Architecture

II. Types and Architectures for CAPCTHA

A. Text based Captchas: These are very simple to implement. These are very effective and need large data of questions as shown in figure 1.1.

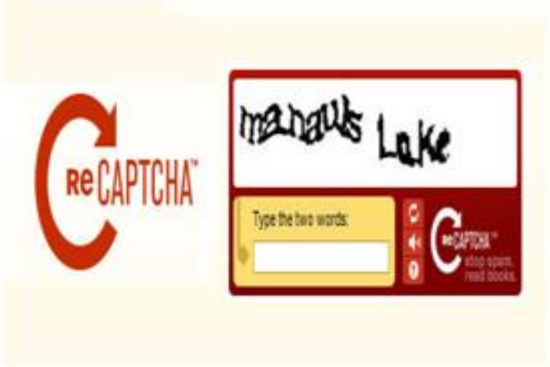


Figure1.1 Captcha

B.Graphics-Based Captchas: These are challenging tests where the user needs to guess the images with some similarity , shown in figure 1.2 and 1.3.

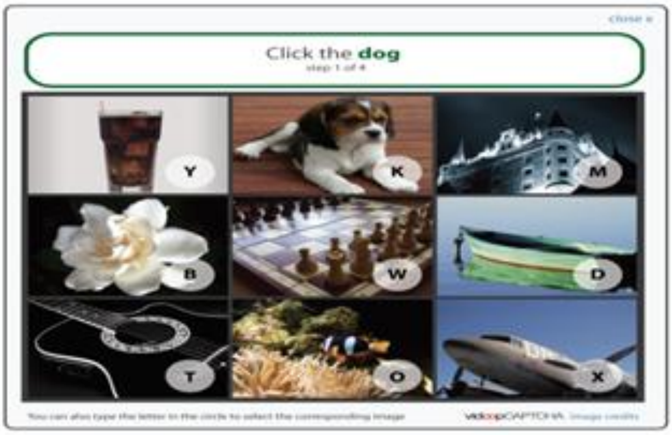
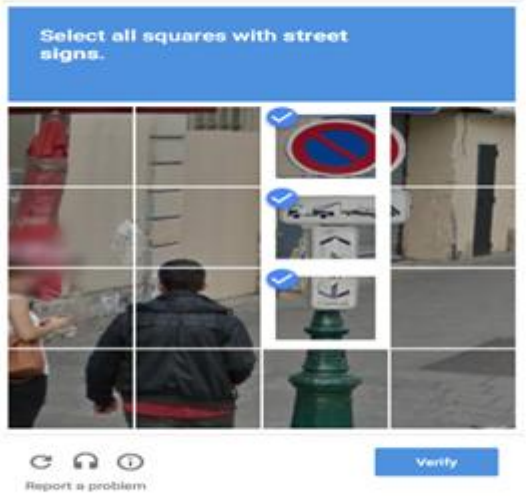

close x
Click the dog
step 1 of 4
Y
K
M
B
W
D
T
O
X
You can also type the letter in the circle to select the corresponding image
vidoopCAPTCHA image credits

Figure 1.2 Guess pictures Captcha



C. Audio-based Captchas: These CAPTCHAs are developed for users with visual disability as shown in figure 1.4. [7].

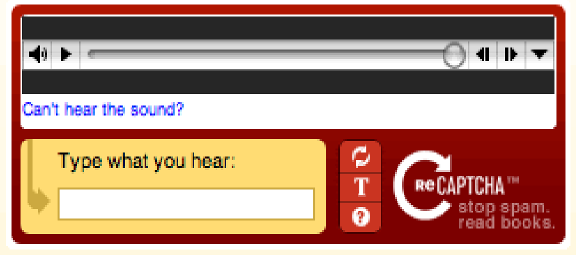


Figure1.4 Captcha for Visual Disables

D. Convolution Neural Network (CNN) Algorithm

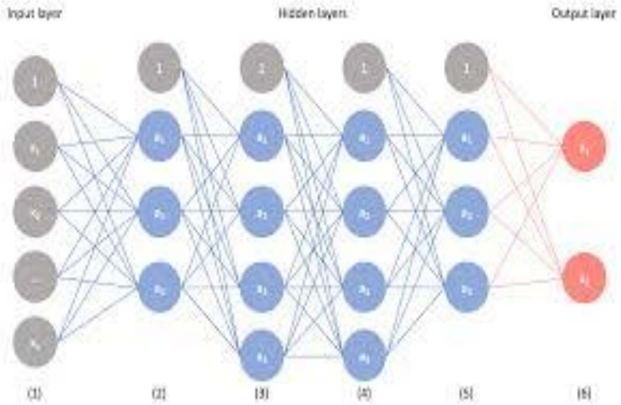


Figure 2.1 CNN) Algorithm

The data is given as input to the model and the model is made in such a way so as to give output of each layer which is known as feed forward. Then we can calculate the error by applying an error function like cross entropy, square loss error and more. Then the derivatives are calculated by back propagating into the model. This is basically applied to get the loss [8]. A python code is used for a neural network which has random inputs and two hidden layers.

1. activation = lambda x: 1.0/(1.0 + np.exp(-x)) # sigmoid function
2. input = np.random.randn(3, 1)

3. hidden_1 = activation(np.dot(W1, input) + b1)
4. hidden_2 = activation(np.dot(W2, hidden_1) + b2)
5. output = np.dot(W3, hidden_2) + b3

where w1,w2,w3,b1,b2,b3 are parameters of the model shown in fig 2.2.

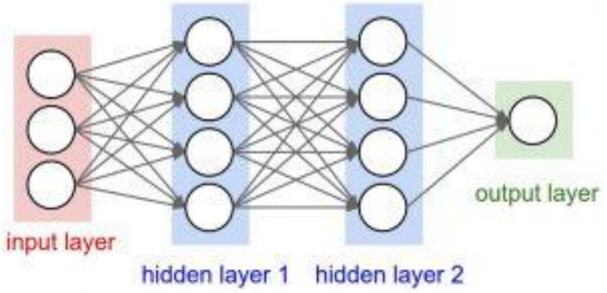


Figure 2.2 CNN Hidden Layers

Procedure

- In order to train a machine, a trained data is needed. To decipher CAPTCHA , we need some training data in this way:

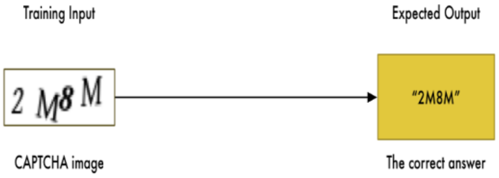


Figure2. 3. Captcha Traning

- Data which is collected from different sources forms a training data, and can be use directly for training a neural network.

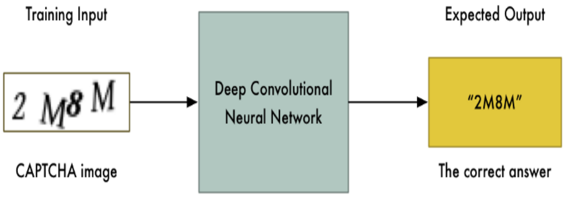


Figure 2. 4 Data Collection

- So we'll start with a raw CAPTCHA image:

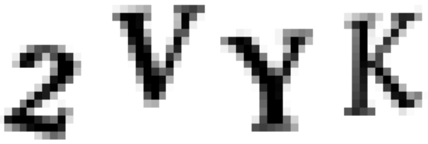


Figure 2.5 Raw Captcha

Convert the Captcha image into clean black and white image, in order to find continuous regions very easily , figure 2.6:

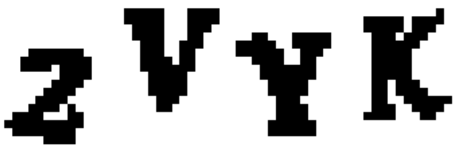


Figure 2.6 Continuous Captcha

For this , use OpenCV's *findContours()* function for detecting separate individual parts of the image which has continuous blobs of pixels of the same color:

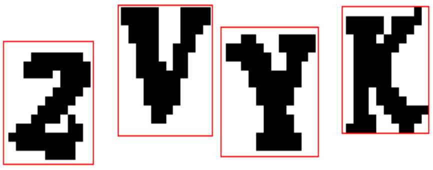


Figure 2.7 Saving Captcha Blocks

Save each image as a separate individual image file.

Sometimes if CAPTCHAs contain overlapping letters as shown below Figure 2.7- 2.10:

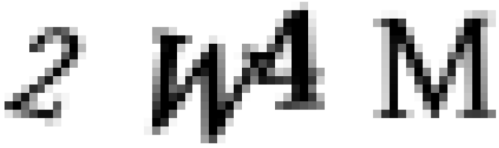


Figure 2.8 Overlapping Captcha

Then find any contour area wider than it is tall, it means that the letters are merged together. When we come across cases like this, we can split the letters exactly in down half middle and divide them as two individual letters.

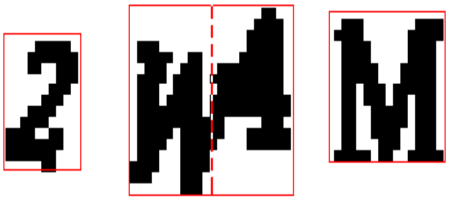


Figure 2.9 Captcha resolving process

- Now apply a procedure to extract individual symbols or letters, which can be run across all the Captcha images which we available. The main goal is to collect various variations of every letter. Save all the individual letters as separate files put together in one folder.

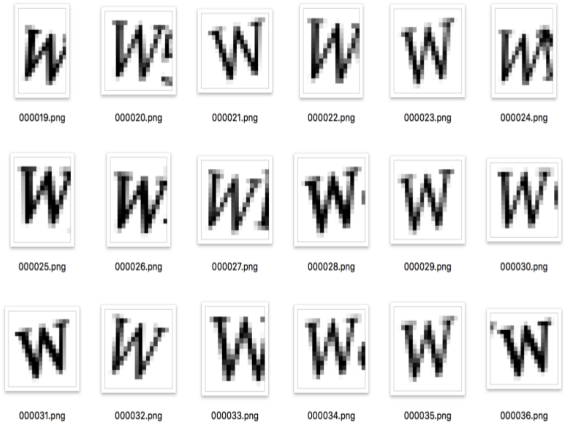


Figure 2.10 Captcha characters resolved

III. Building and Training the Neural Network

- Here we use a simple convolutional neural network architecture with two convolutional layers and two fully-connected layers. Defining this neural network architecture takes a few lines of code using Keras[9] as shown in figure 3.1.

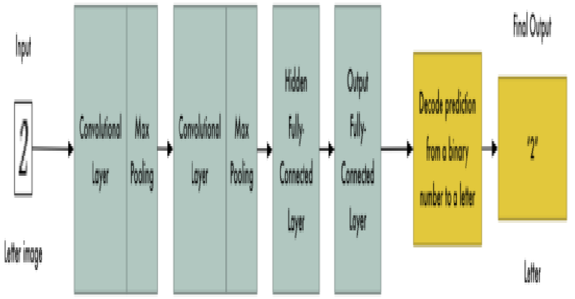

Input
2
Letter image
Convolutional Layer
Max Pooling
Convolutional Layer
Max Pooling
Hidden Fully-Connected Layer
Output Fully-Connected Layer
Decode prediction from a binary number to a letter
Final Output
"2"
Letter

Figure 4.1 Generic Training model

- By passing 10 inputs to the trained model, it can hit almost 100% accuracy. By this we must automatically bypass the above Captcha whenever needed[8].

IV. Use Trained Model to Decipher Captchas

- As the trained neural network is available , it makes the task of breaking real Captchas simple.
- Take an image of CAPTCHA from any webpage. Classify the image into their separate respective letters as followed while training a dataset. Now input the letters individually to the Neural network model for predicting each letter.
- The letters which are being predicted separately are together considered to be as the answer to the given Captcha image i.e. figure-4.2.

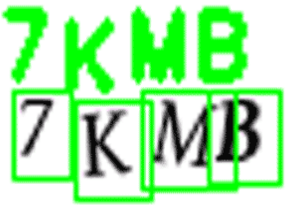


Figure 4.2 Captcha Image

V. Merits and Demerits

Advantages

- As it is completely software involved, human work is being reduced.
- This model provides accurate results.
- It consumes less time to decipher many such Captchas.

Disadvantages

- There may be rare possibilities of difficulty in reading Captcha.
- It can be misused in any way if there are no preventive measures.
- Sometimes consumes more time to decipher[10].

V. Conclusion

For the growing need of the present and future generations, Captcha human authentication was successful on the internet for decades. More and more technical advancements can have many more advanced neural network architectures like capsule networks which has very less steps and is more general and can be applied to various schemes of Captchas without modifying them.

References :

[1] L. von Ahn, M. Blum and J. Langford. "Telling Humans and Computer Apart Automatically", CACM, vol. 47, 2004.

[2] L. von Ahn, M. Blum, N. J. Hopper and J. Langford, "CAPTCHA: Using hard AI problems for security , Proceedings of te 22nd international conference on theory and applications of cryptographics techniques,2003.

[3] "The reCAPTCHA Project – Carnegie Mellon-University CyLab". www.cylab.cmu.edu.

[4] Von Ahn, Luis; Blum, Manuel; Hopper, Nicholas J.; Langford, John (May 2003). CAPTCHA: Using Hard AI Problems for Security. EUROCRYPT 2003: International Conference on the Theory and Applications of Cryptographic Techniques.

[5] May, Matt (2005-11-23). "Inaccessibility of CAPTCHA". W3C. Retrieved 2015-04-27.

[6] Bursztein, E., Bethard, S., Fabry, C., Mitchell, J. C., & Jurafsky, D. (n.d.). Retrieved March 30, 2018,

[7] Jump up to "idrive turing patent application". Retrieved 2017-05-19.

[8] "h2g2 – An Explanation of l33t Speak – Edited Entry". Retrieved 2015-06-03.

[9] "idrive turing signup page". Retrieved 2017-05-19.

[10] Stringham, Edward P (2015). Private Governance : creating order in economic and social life. Oxford University Press. p. 105. ISBN 978-0-19-936516-6.