International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

AUTOMATED CAPTCHA RECOGNITION WITH CONVOLUTIONAL NEURAL NETWORKS

[1]L.Priyanka, [2]Gangula Rajkumar
[1]Assistant Professor, Department of MCA Student, Sree Chaitanya College of Engineering, Karimnagar

[2]MCA Student, Department of MCA Student, Sree Chaitanya College of Engineering, Karimnagar

ABSTRACT

Automated tests called CAPTCHAs are made to differentiate between machines and people. They prohibit programs from abusing online services and using internet resources because they are difficult for machines to solve but simple for people. Convolutional neural networks might be used to answer the CAPTCHA tests automatically and effectively. The high-accuracy CAPTCHA identification methods used nowadays can be physically complex. Our team thus investigated an alternative method for solving CAPTCHAs using a Convolutional Neural Network, which is more effective in terms of run time and structural complexity, with the potential to improve accuracy through image processing. Additionally, we used CAPTCHA datasets with background noise and character adhesion to evaluate our networks.

1. INTRODUCTION

CAPTCHA (Completely Automated Public Turing Test to tell Computers and Humans Apart) is an automated test created to prevent websites from being repeatedly accessed by an automatic program in a short period of time and wasting network resources. Most service providers online have implemented CAPTCHA tests before the user is allowed to commit certain actions, such as submitting a form. Among all the CAPTCHAs, commonly used types contain low resolution, deformed characters with character adhesions and background noise, which the user must read and type correctly into an input box. This is a relatively simple task for humans, taking an average of 10 seconds to solve [1], but it presents a difficulty for computers, because such noise makes it difficult for a program to differentiate the characters from them. However, using Convolutional Neural Networks (CNN), these CAPTCHA tests could be solved efficiently and accurately by computers. Creating a simple, efficient and accurate method to recognize CAPTCHAs can assist with the verification of the security of existing forms of CAPTCHAs and the creation of new, more secure ones. The same approach for CAPTCHA recognition could also be applied in several other fields, including handwriting recognition, license plate recognition, and many more. The focus of our research is to optimize a CNN model for structural simplicity, accuracy, and amount of training data required. To further improve the accuracy of the models and decrease the training data needed, we also experimented with preprocessing the CAPTCHA images using methods such as Fourier Transform

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

with the goal of eliminating background noise.

2. LITERATURAL SURVEY

TITLE: How good are humans at solving CAPTCHAs

AUTHOR: Bursztein, E., Bethard, S., Mitchell, J.C., Jurafsky, D., Fabry, C.

Captchas are designed to be easy for humans but hard for machines. However, most recent research has focused only on making them hard for machines. In this paper, we present what is to the best of our knowledge the first large scale evaluation of captchas from the human perspective, with the goal of assessing how much friction captchas present to the average user. For the purpose of this study we have asked workers from Amazon's Mechanical Turk and an underground captchabreaking service to solve more than 318 000 captchas issued from the 21 most popular captcha schemes (13 images schemes and 8 audio scheme). Analysis of the resulting data reveals that captchas are often difficult for humans, with audio captchas being particularly problematic. We also find some demographic trends indicating, for example, that non-native speakers of English are slower in general and less accurate on English-centric captcha schemes. Evidence from a week's worth of eBay captchas (14,000,000 samples) suggests that the solving accuracies found in our study are close to real-world values, and that improving audio captchas should become a priority, as nearly 1% of all captchas are delivered as audio rather than images. Finally our study also reveals that it is more effective for an attacker to use Mechanical Turk to solve captchas than an underground service.

TITLE: **Gradient-based learning applied to document recognition**

AUTHOR: Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.

Multilayer neural networks trained with the back-propagation algorithm constitute the best example of a successful gradient based learning technique. Given an appropriate network architecture, gradient-based learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns, such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task. Convolutional neural networks, which are specifically designed to deal with the variability of 2D shapes, are shown to outperform all other techniques. Real-life document recognition systems are composed of multiple modules including field extraction, segmentation recognition, and language modeling. A new learning paradigm, called graph transformer networks (GTN), allows such multimodule systems to be trained globally using gradient-based methods so as to minimize an overall performance measure. Two systems for online handwriting recognition are described. Experiments demonstrate the advantage of global training, and the flexibility of graph transformer networks. A graph transformer network for reading a bank cheque is also described. It uses

International Journal For Advanced Research
In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

convolutional neural network character recognizers combined with global training techniques to provide record accuracy on business and personal cheques. It is deployed commercially and reads several million cheques per day.

3. EXISTING SYSTEM

Most service providers online have implemented CAPTCHA tests before the user is allowed to commit certain actions, such as submitting a form. Among all the CAPTCHAs, commonly used types contain low resolution, deformed characters with character adhesions and background noise, which the user must read and type correctly into an input box. This is a relatively simple task for humans, taking an average of 10 seconds to solve [1], but it presents a difficulty for computers, because such noise makes it difficult for a program to differentiate the characters from them.

Disadvantage

1. Many kinds of CAPTCHAs, text-based CAPTCHAs are commonly used even if they are not the most secure option because they are cheap, convenient, and user friendly.
2. Because text-based CAPTCHAs are more vulnerable to attacks and less secure than originally intended, improvements need to be made.

3. PROPOSED SYSTEM

Using Convolutional Neural Networks (CNN), these CAPTCHA tests could be solved efficiently and accurately by computers. Creating a simple, efficient and accurate method to recognize CAPTCHAs can assist with the verification of the security of existing forms of CAPTCHAs and the creation of new, more secure ones. The same approach for CAPTCHA recognition could also be applied in several other fields, including handwriting recognition, license plate recognition, and many more.

Advantages

1. CNN is an efficient and accurate method of recognizing CAPTCHAs
2. Improve the security of text-based CAPTCHAs

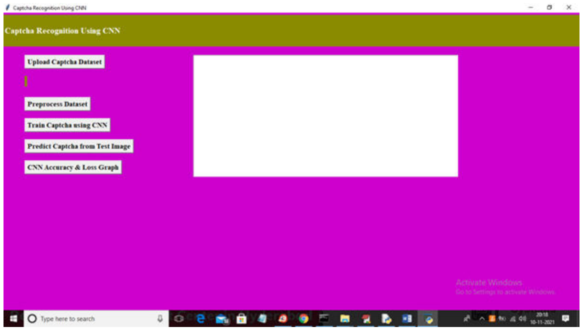
4. IMPLEMENTATION

1) Data Collection: Using this module we will upload CAPTCHA image dataset to application
2) Data Preprocessing: using this module we will read all CAPTCHA image and after applying Preprocessing we will extract features from all reviews.
3) Train CNN Algorithm: we are trainging the cnn algorithm with the dataset
4) Predict Captcha: we are predicting the CAPTCHA by giving the test images

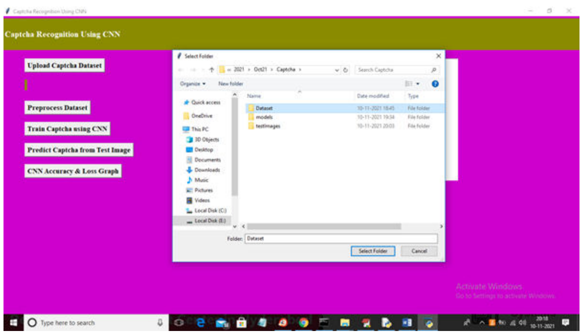
5. SCREEN SHOTS

To run project double click on ‘run.bat’ file to get below screen

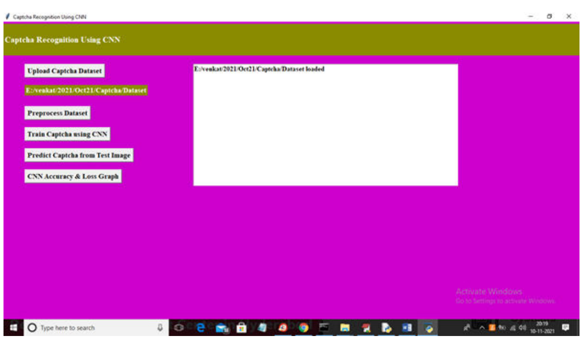
International Journal For Advanced Research In Science & Technology
IJARST
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362



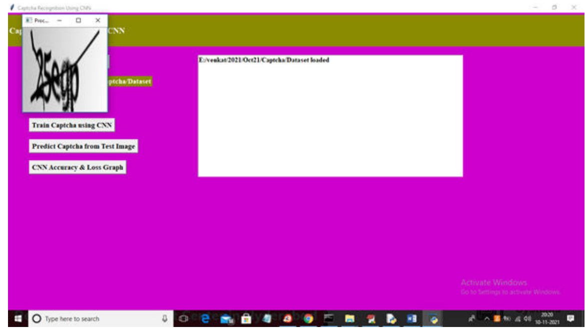
In above screen click on 'Upload Captcha Dataset' button to upload dataset and to get below screen



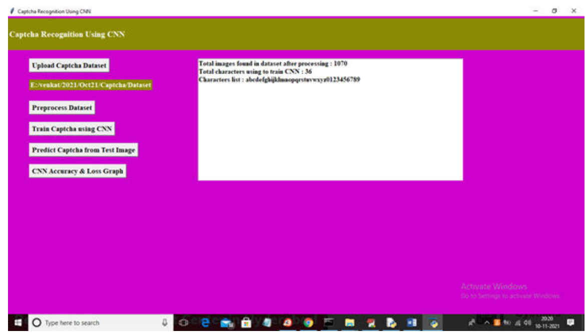
In above screen selecting and uploading 'Dataset' folder and then click on 'Select Folder' button to load dataset and to get below screen



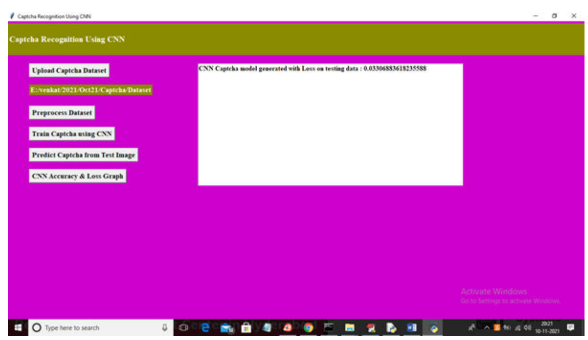
In above screen dataset loaded and now click on 'Preprocess Dataset' button to process all images and to get below screen



In above screen displaying sample grey scale normalized image and then close above image to get below output

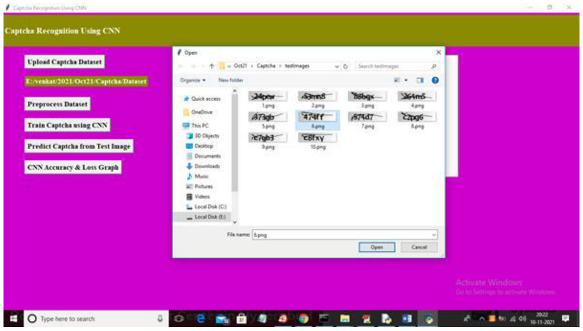


In above screen we can see total dataset images and characters used to train CNN model. Now dataset is ready and now click on 'Train Captcha using CNN' button to train CNN model and calculate loss value

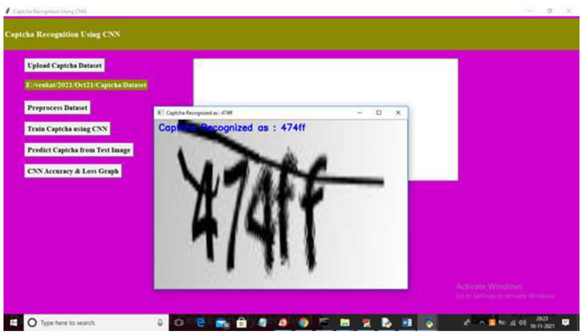


In above screen we got CNN loss value as 0.033 so accuracy will be 100 – 0.033 = 99.967 and now model is ready and now click on 'Predict Captcha from Test Image' button to upload test image like below screen

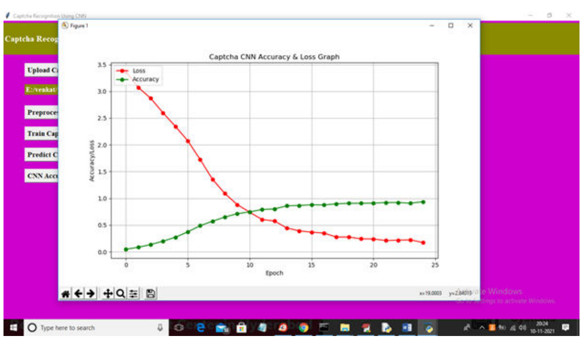
IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362



In above screen selecting and uploading '6.png' image and then click on 'Open' button to get below output



In above screen Captcha is recognized as '474ff' and similarly you can upload other images and test them. Now close above image and then click on 'CNN Accuracy & Loss Graph' button to get CNN training performance



In above graph x-axis represents EPOCH and y-axis represents accuracy/loss values and in above graph red line represents LOSS and green line represents accuracy and in above graph we can see with each increasing epoch accuracy value got increased and loss values got decreased which indicates CNN trained accurately on dataset.

6. CONCLUSION

Although there are many other types of CAPTCHAs, text-based CAPTCHAs are frequently utilised because they are affordable, practical, and easy to use, even if they are not the most secure option. Improvements are required because text-based CAPTCHAs are less secure than anticipated and more susceptible to assaults. By identifying their flaws, text-based CAPTCHAs may be made more secure by developing more accurate and efficient methods of solving them. All things considered, CNN is a reliable and effective way to identify CAPTCHAs, and further advancements in its application might increase the security of text-based CAPTCHAs. Our team built three CNN networks that are structurally more efficient than many of the current methods of high accuracy CAPTCHA recognition, and tested them on three different CAPTCHA datasets to see if they could be used to accurately recognise CAPTCHAs with character adhesions and background noise. CNN networks have shorter runtimes and less structural complexity than other methods. With just 1070 samples chosen for training from each dataset, the findings demonstrate that these networks can achieve high recognition accuracy while having less structural complexity, with Network 1 achieving 94.67% accuracy on the first dataset. We think that further training might significantly increase accuracy, even in cases when it is poor. According to these findings, the third dataset—which consists of ten randomly chosen digits and all 26 capital and lowercase letters—was significantly more difficult for the models to identify than the other two datasets, requiring more training to achieve the required accuracy. Character adhesions, slant, and background dots and lines also served as distractions. As part of our continuous efforts, the future security of text-based CAPTCHAs is still an unresolved issue.

REFERENCES

[1] Bursztein, E., Bethard, S., Mitchell, J.C., Jurafsky, D., Fabry, C. (2010) How good are humans at solving CAPTCHAs? a large scale evaluation. IEEE S&P ’10.

[2] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE.

[3] Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V. (2014) Multi-digit number recognition from street view imagery using deep convolutional neural networks. ICLR.

[4] Zhang, L., Huang, S.G., Shi, Z.X., Hu, R.G. (2011) CAPTCHA recognition method based on LSTM RNN. Pattern Recogn., 1: 40–47.

[5] Yan, J., Ahmad, A.S.E. (2008) A low-cost attack on a Microsoft CAPTCHA. Proceedings of the ACM Conference on Computer and Communications Security, 543–554.

[6] Stark, F., Hazirbaş, C., Triebel, R., Cremers, D. (2015) Captcha recognition with active deep learning. GCPR Workshop on New Challenges in Neural Computation.

[7] Wang, J., Qin, J.H., Xiang, X.Y., Tan, Y., Pan, N. (2019) CAPTCHA recognition based on deep convolutional neural network. Math. Biosci. Eng., 16: 5851–5861.

[8] Kaggle. (2018) CAPTCHA Images. https://www.kaggle.com/fournierp/captcha-version-2-images.

[9] Kaggle. (2020) CaptchaImages1070. https://www.kaggle.com/datascientistsohail/captchaimages1070.