Cracking CAPTCHA using Deep Learning

Team Members: Dhruv Kaushik-MT18037, Subhani Shaik-MT18117,Gurpreet Singh-MT18098

I.MOTIVATION AND PROBELM STATEMENT

A.Motivation

A CAPTCHA(Completely Automated Public Turing test to tell Computers and Human Apart), now a days which is commonly used in internet for security purpose. But due to advanced learning techniques these CAPTCHAs are often decoded easily. So, we are interested in testing the sensitivity and strength of the CAPTCHA images.

B.Problem Statement

We are having CAPTCHA images and we need to depict what is value of the captcha that is written in the image.

II. LITERATURE REVIEW

1. Yashwanth Chanamolu (2009), “An Orientation Based Image CAPTCHA thesis paper”.
2. Matous Pistora (2016), “Pattern Recognition of CAPTCHA research paper”.
3. Tan,Steinbach and Kumar, “Introduction to Data Mining(Textbook)”.
4. Walid Khalifa Abdullah Hasan, “A Survey of Current Research on CAPTCHA”.
5. Documentation on svc.SVM from Sklearn.
6. Understanding Machine Learning : From Theory to Algorithm by Shalev Shwartz.
7. Keras documentation for convolution neural network.

III. DATASET AND PREPROCESSING

For our problem statement we used single letter and four letter captchas.So, we generated the CAPTCHA images using python library-‘captcha’. By using this library we generated images with random noise by setting specific dimension.

A. Grayscale conversion and Denoising image

Initially we removed the overlapping characters from the dataset and we converted the images into grayscale images. These grayscale images are denoised usng median filter with corresponding parameters which are perfect for image without losing the data from images.

B.Segmentation

The denoised images are then segmented using Watershed algorithm which highlights the character(object) in our problem. We applied Grid based clustering to crop the images into individual characters by using some specific dimensions.After denoising and segmenting image look like:

IV. TECHNIQUES

We used supervised learning techniques like SVM ,CNN models and we used Unsupervised learning technique – Kmeans by taking k=26 clusters.
In these Kmeans and SVM are baseline models and CNN is our advanced model.

K-Means clustering:

The Kmean clustering technique is used to group similar data items into clusters.we created 26 clusters for our training dataset such that each cluster contains datapoints corresponding to certain character.

For each cluster, a representative is found in the kmeans clustering.Now when a test point arrives, the model trained by using kmeans clustering tries to map the test datapoint to the nearest representative based on its similarity with that representative calculated by using Eculidian distance.

SVM(Support Vector Machines):

SVM are the classifiers that can create a boundary between the datapoints of different classes. So, that the points can be uniquely identified as the class to which they belong.

We used Linear SVM and trained it using 20000 single letter (each character are approximately 900)captcha images. We used Linear SVM because the data is linearly seperable. The SVM will generally fails in the cases when two different class datapoints coincide in their feature vector that is when the feature vector of the points which belong to different classes are moreover the same.In these kind of overlapping cases, we can not make a decision boundary. So, SVM fails in these kind of scenarios.

Before feeding the inputs to SVM, the data's feature vector is done linear from 2D to 1D.

CNN(Convolution Neural Network):

The network tries to learn different patterns in the images and classify the images with same pattern to a single class.

The CNN model was trained on single letter captcha images and then it was used to classify the captcha having four letters.So, in this multiple character captcha image, it is segmented into four images and the prediction is done on each image. After that all four images will be compared to the ground truth value of the image.

Our CNN architecture is as follows:

It has 4 Convolution layers,4 Maxpool layers, 2 Dense layers and 1 output layer which is 'softmax' activation. It also fails in the same scenario as mentioned in SVM due to same reason.

V. RESULTS

Accuracy

Model	train_data accuracy	test_data accuracy
Kmeans(Single letter)	-	0.20
SVM(Single letter)	1.0	0.996
CNN(Single letter)	0.998	0.997
CNN(Four letters)	-	0.816

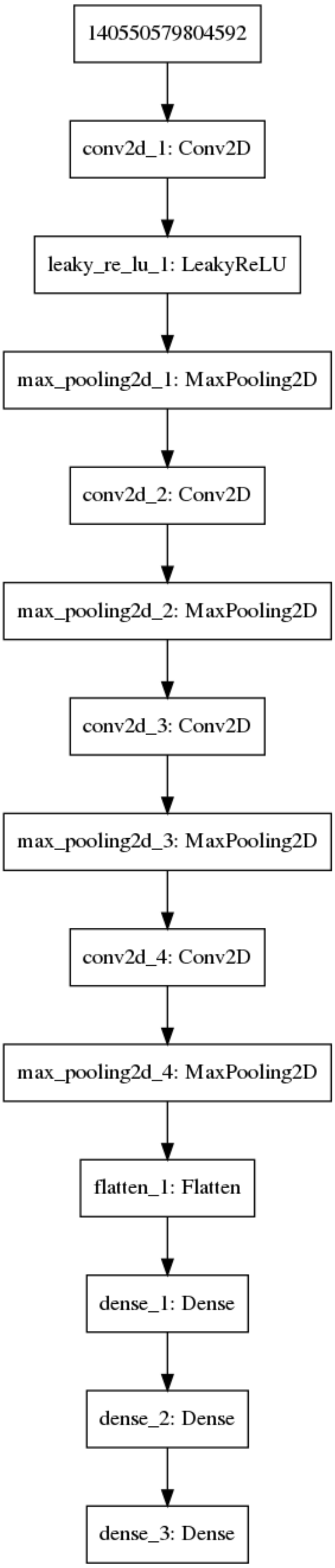

Fig : Architecture of CNN

Model	f1_score
Kmeans(Single letter)	-
SVM(Single letter)	0.9959(macro),0.9961 (micro)
CNN(Four letters)	0.69019(macro),0.8167(micro)

VI. ANALYSIS

SVM and CNN:

Assuming that our data is well diversed , the model is trained very well. They are not underfitting and overfitting. The accuracy is mentioned above prove this. The accuracy is only fails during the overlapping of features of datapoints.

Kmeans clustering:

Because of it is clustering technique, though we tried to use it for our classification problem.It worked very poorly in predicting for any given datapoint.

VII.SUMMARY

In our learning model, the CNN model is best.Due to the fact that it works on different kinds of patterns for which the model is more generalized.So, if there is any image that is outside of the dataset from which we created then in that case it has high probability to classify it correctly. Whereas SVM does not recognize patterns,so it may not work well than CNN.

VIII. CONCLUSION

Among our model, the CNN technique is more generalized and not bias.The only problem is it need high computation to train the dataset.In the case of CAPTCHAs we end with the conclusion that if our training data ensures that all the captchas would come out of similar distribution of our training data then we can train the model and detect them very accurately.This itself is main challenge in cracking the captchas because there are possible many patterns and that would take lot of computation and work to explore and create dataset from all those distributions and train the model. Even then also you cannot ensure all cases of CAPTCHAs.