reCAPTCHA

by Dominic Grazioli, Johann Ryan, Indresh Srivastava, Rithvik Subramanya

Introduction

Project Goal: Identify the text within reCAPTCHA images

Dataset:

Easy Dataset: https://www.kaggle.com/genesis16/captcha-4-letter
Hard Dataset: https://www.kaggle.com/fournierp/captcha-version-2-images

Easy Dataset

Hard Dataset

Uses:

Identifying text within old manuscripts
Paper: Ahn, L. von, Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008, September 12). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Retrieved from https://science.sciencemag.org/content/321/5895/1465

Testing security

Github Link: https://github.com/srivasis/reCAPTCHA

Approach to Solving the Problem

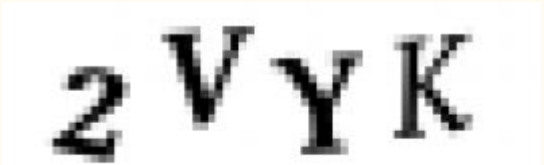


Figure 2.5 Raw Captcha

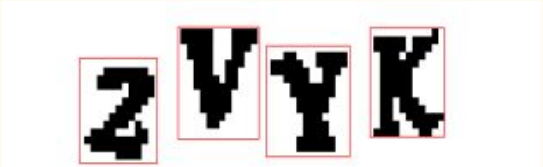


Figure 2.7 Saving Captcha Blocks

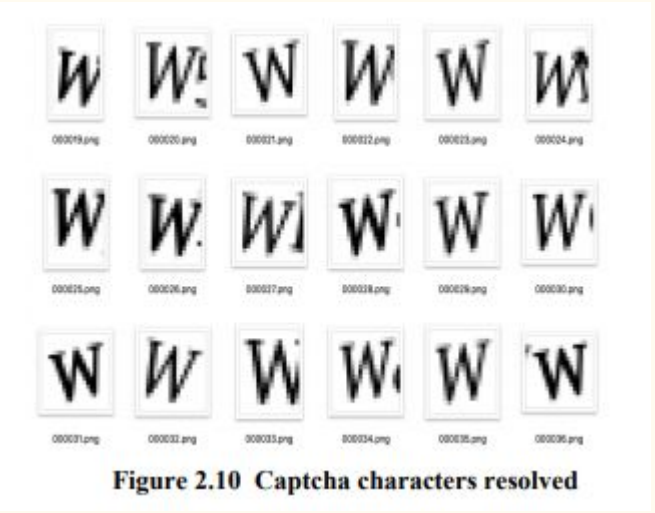
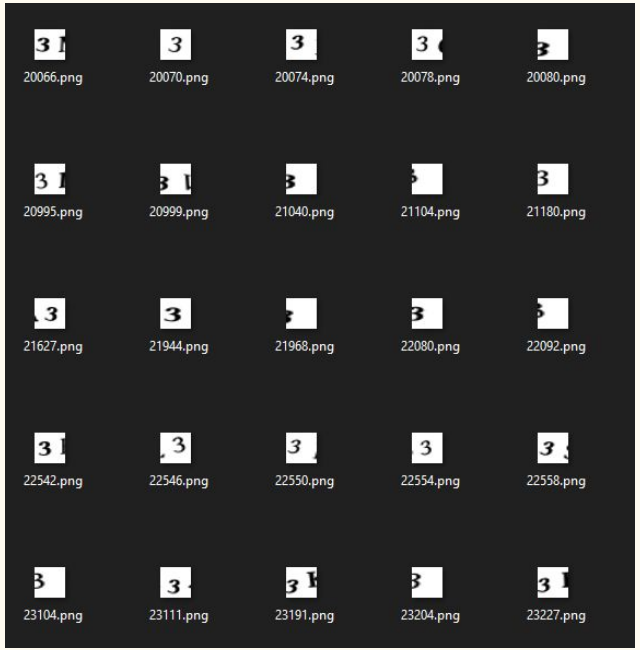


Figure 2.10 Captcha characters resolved

Paper: Dr. T. Venkat Narayana Rao and A. Sai laksmi. Decoding Captcha using Machine Learning for Identifying Criminal Attempts. International Journal on Future Revolution in Computer Science & Communication Engineering, ISSN: 2454-4248, Volume: 5 Issue: 1, pg. 61 – 65.

http://www.ijfrcsce.org/download/browse/Volume_5/January_19_Volume_5_Issue_1/1547886477_19-01-2019.pdf

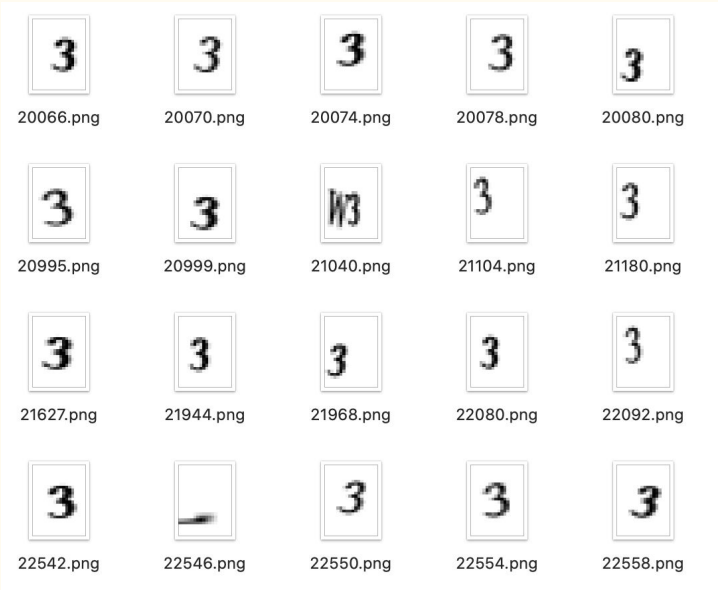
K-Means Character Separation



1. Obtain an image mask of the reCaptcha for pixels > 120
2. Run a K-means clustering on the *x-coordinates only* where $k = 4$
3. Use the midpoints in between the adjacent centroids to split the characters

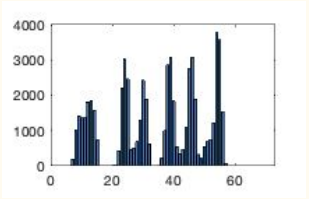
Captures the desired characters for the most part, but many characters are cut off or out of frame

Local Minima Character Separation



1. Columns values subtracted from 255 are summed up

   Example: 

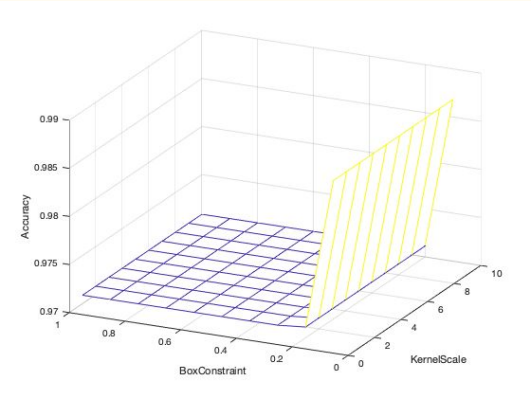
   

2. X-coordinates of the 3 smallest “local minima” are used to split the characters

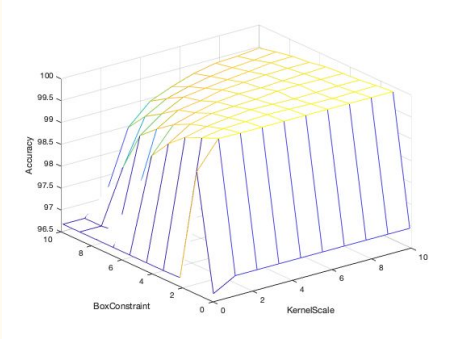
Significantly fewer cut offs than K-means, but fails when characters are too close

Support Vector Machine

- Characters are separated and resized
- HOG Features are taken of each separated character
- Data is split 60/20/20 on characters for training/validation/testing
- A Support Vector Machine is trained for each character
- Kernel Scale and Box Constraint are chosen by hyperparameterization



K-Means
Kernel Scale: 1
Box Constraint: 0.1



Local Minima
Kernel Scale: 1.2
Box Constraint: 1.2

Support Vector Machine Results

K-Means Separation:

Average Validation Accuracy per character:
98.78%

Average Validation Accuracy per reCaptcha:
95.22%

Average Testing Accuracy per character:
95.25%

Average Testing Accuracy per reCaptcha:
82.31%

Local Minima Separation:

Average Validation Accuracy per character:
97.08%

Average Validation Accuracy per reCaptcha:
88.82%

Average Testing Accuracy per character:
96.67%

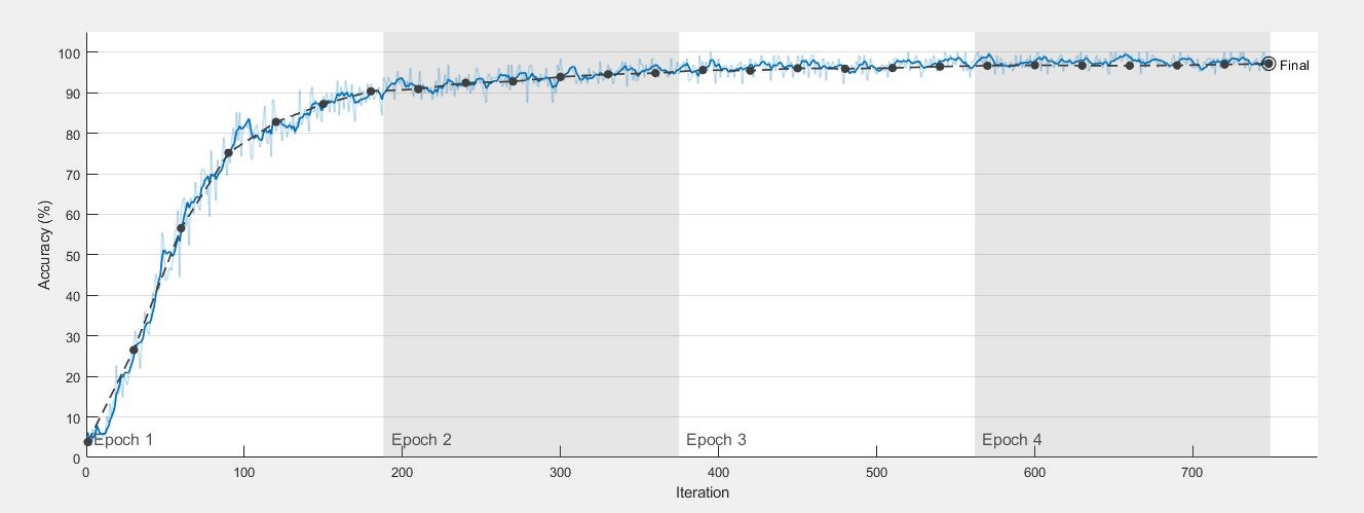
Average Testing Accuracy per reCaptcha:
87.33%

Convolutional Neural Network (CNN)

- Same character and data separations
- Very slightly better than the SVM
  - You also don't need 32 separate CNNs
- Architecture: 3 sets of convolution + ReLU, then the fully connected layer and classification.

```
layers = [
    imageInputLayer([28 28 1])

    convolution2dLayer(3,8,'Padding','same')
    batchNormalizationLayer
    reluLayer

    maxPooling2dLayer(2,'Stride',2)

    convolution2dLayer(3,16,'Padding','same')
    batchNormalizationLayer
    reluLayer

    maxPooling2dLayer(2,'Stride',2)

    convolution2dLayer(3,32,'Padding','same')
    batchNormalizationLayer
    reluLayer

    fullyConnectedLayer(32)
    softmaxLayer
    classificationLayer];
```

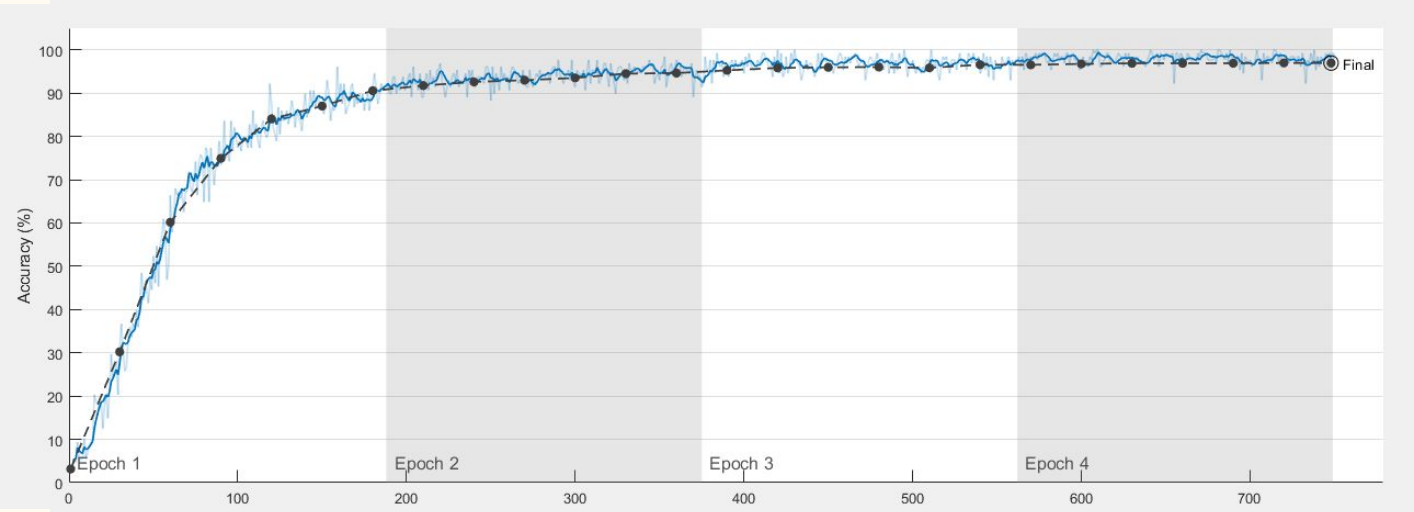
CNN Results



K-Means (Validation):
96.96% Accuracy per Character
88.38% Average Accuracy per reCaptcha

K-Means (Testing):
96.44% Accuracy per Character
86.50% Average Accuracy per reCaptcha

CNN Results



Local Minima (Validation):
96.99% Accuracy per Character
88.49% Average Accuracy per reCaptcha

Local Minima (Testing):
96.91% Accuracy per Character
88.20% Average Accuracy per reCaptcha

Final Results

SVM (K-Means separation) Testing Accuracy per reCaptcha: 82.31%

SVM (Local Minima separation) Testing Accuracy per reCaptcha: 87.33%

CNN (K-Means separation) Testing Accuracy per reCaptcha: 86.50%

CNN (Local Minima separation) Testing Accuracy per reCaptcha: 88.20%