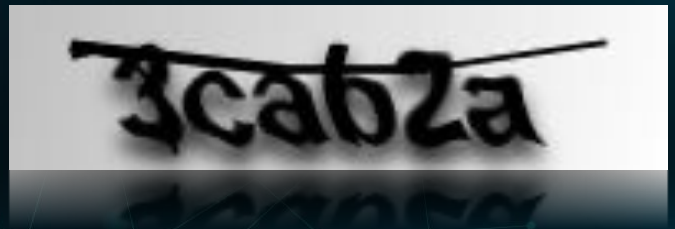
CAPTCHA Images Decoder



Text-based CAPTCHA

Group Members:

- Yuxi Chen
- Zijie Tan
- Lijiangnan Tian
- Ze Hui Peng

Voice-over: Ze Hui Peng

CONTENTS

Introduction

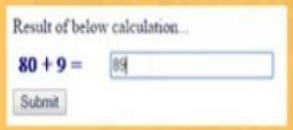
- CAPTCHA (Completely Automated Public Turing test to tell Computers and Human Apart) has been a popular method to distinguish robots from human users
- Our goal is to use different algorithms to decode the CAPTCHA images and compare their performance

Evolution of CAPTCHA

Text-based Captcha

ReCAPTCHA

3D Captcha

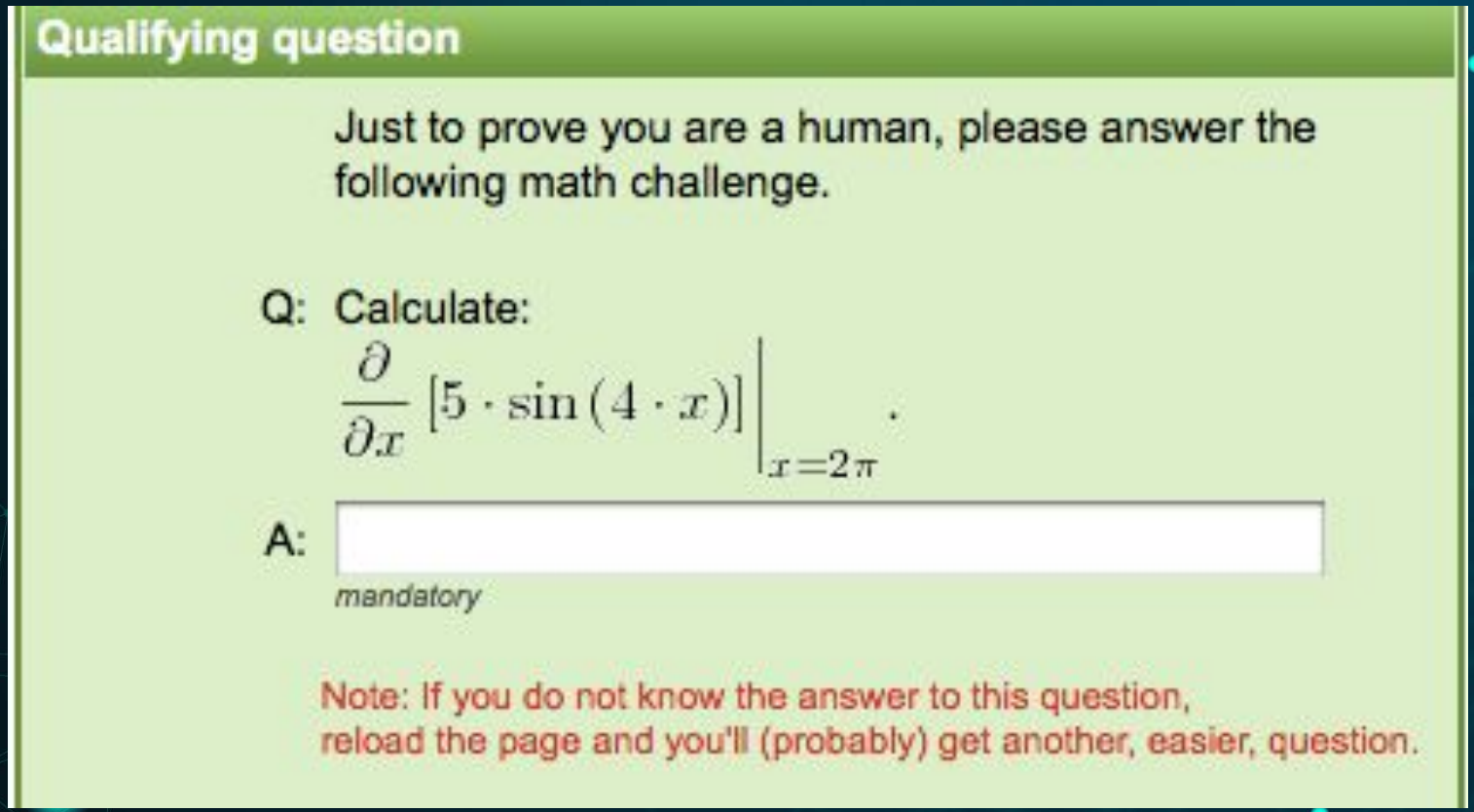


Mathematical Captcha

Image-based Captcha

Source: https://www.youtube.com/watch?v=YxKYrAjweGU

Evolution of CAPTCHA



Source: https://www.theregister.com/2011/03/09/calculus_based_captcha/

Data

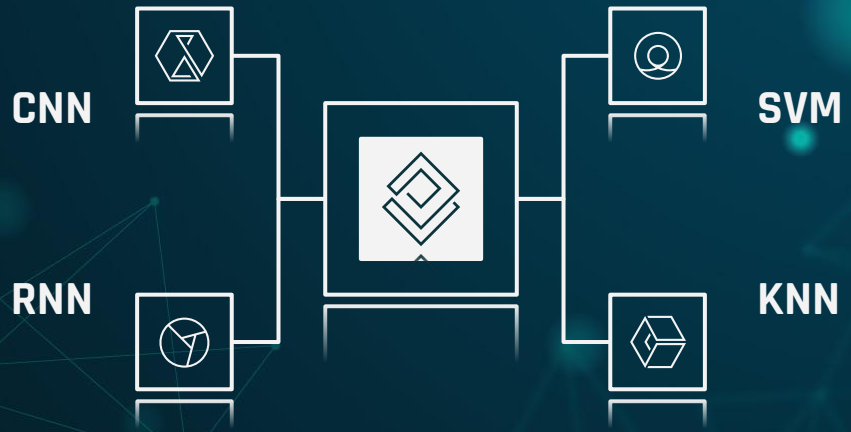
- 10,000 CAPTCHA images of size 200 × 50 pixels[1]
  - Each has darkened background & 6-character-long string
  - Including 2 type of noise: shadow and fisheye effect
- Generated by the `Google Kaptcha Library`[2][3]
  - In the name `{content}_{index}.jpg`
- Supervised classification
  - Multiclass Labels (36 classes, including 10 digits & 26 letters)

A sample CAPTCHA image with content `882m62`

Methodology

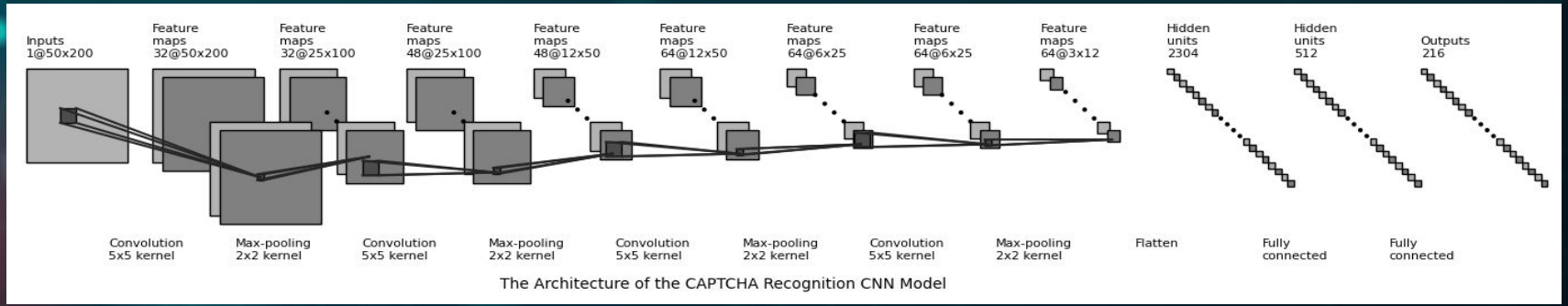
- Segmentation-free models
  - Convolutional Neural Network (CNN)
  - Recurrent Neural Network (RNN)
- Segmentation-based models
  - Support Vector Machine (SVM)
  - k-Nearest Neighbours (KNN)
- Measurement
  - Performance is measured by accuracy
  - Correctly predict all characters in a CAPTCHA?

Segmentation-free VS Segmentation-based



Segmentation-free method: CNN

- **Convolutional Neural Network**
  - Reduce large data volume images to small data volume
  - Retains image characteristics
- **Model**
  - ReLU activation function
  - 4 Convolution Layers, 4 Max-pooling Layers, 2 Fully-connected Layers
  - Outputs a sequence of length 216 (6×36)

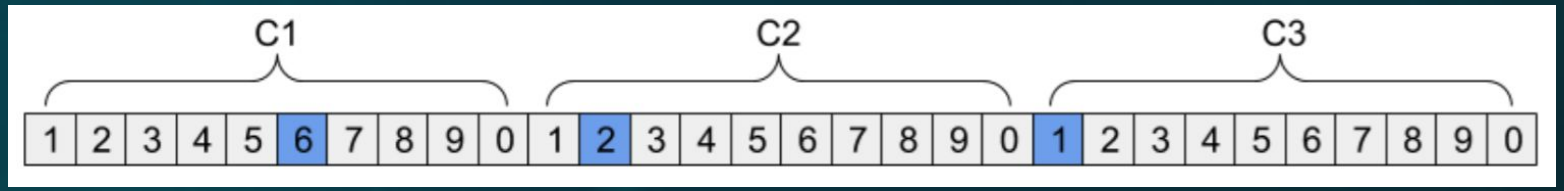


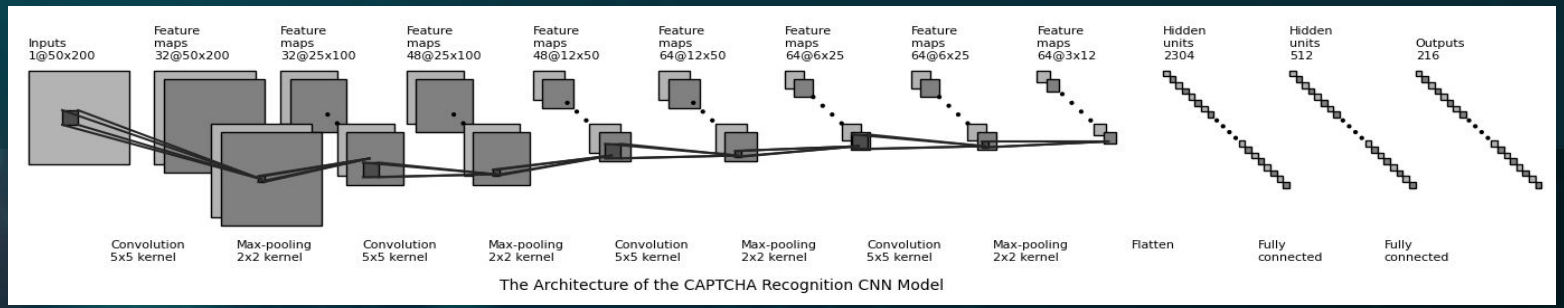
The Architecture of the CAPTCHA Recognition CNN Model

Segmentation-free method: CNN

Flow of Data

- The model receives image data as input
- Outputs a sequence of length 216 (6×36)
- Sequence divided into 6 equal-sized segments (36, the same as the number of labels)
- Obtain the index of the label according to the index of the *maximum* value in segments





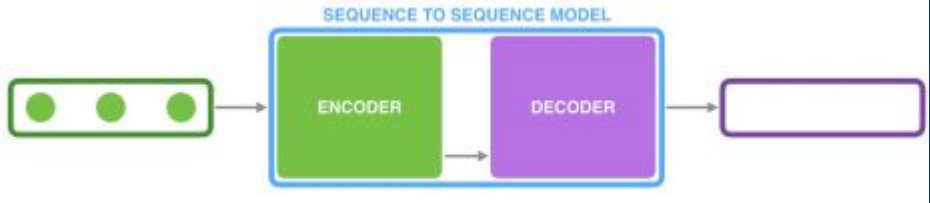
The Architecture of the CAPTCHA Recognition CNN Model

Segmentation-free method: RNN

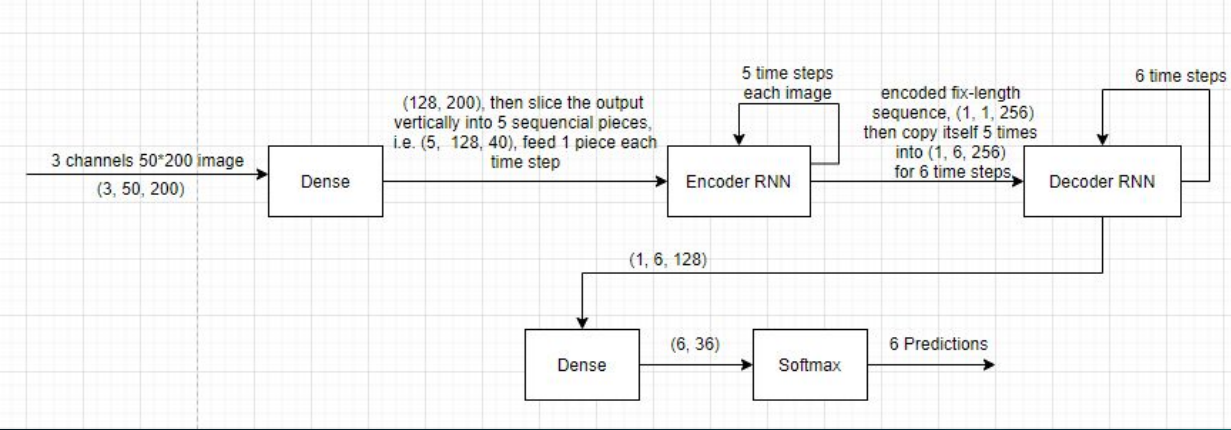
- Recurrent Neural Network
  - Potential to solve *variable-length* CAPTCHAs
- However
  - Traditional RNN has poor performance on such an Image Recognition problem
    - Accuracy 0.02% after 100 epochs
  - Variation: Encoder-Decoder RNN

Segmentation-free methods: RNN

- Encoder-Decoder Model (seq2seq)

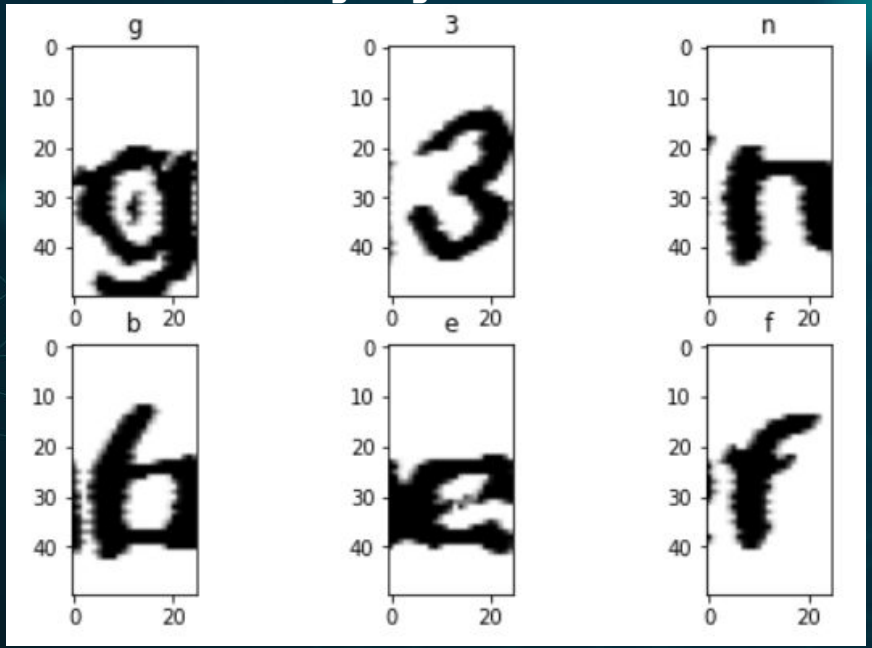


- Flow of data



Flow of Data for SVM and KNN

Image segmentation

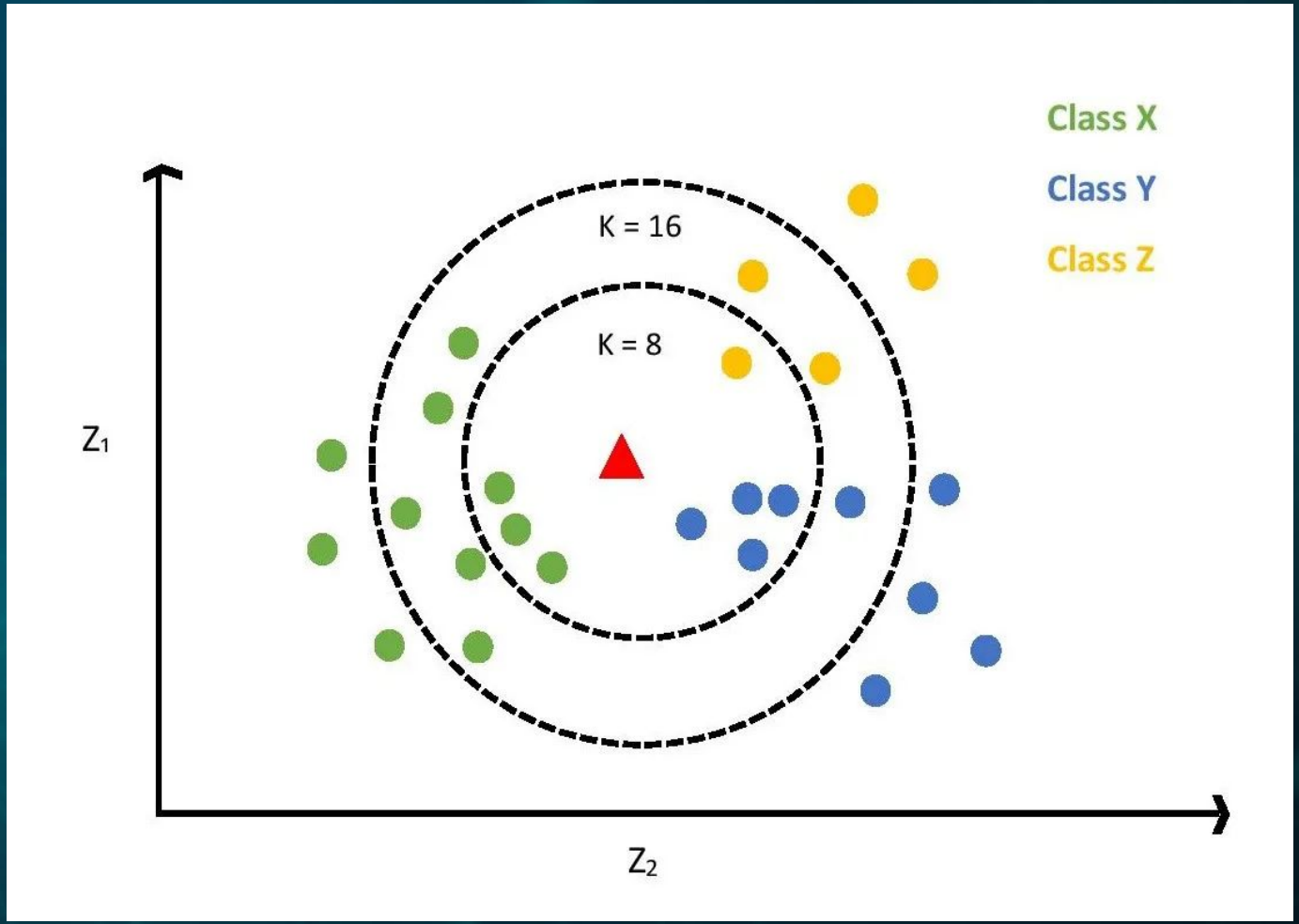


Segmentation-based methods: SVM

- Support Vector Machine
- Tool used: `sklearn.svm.SVC`
- *Segmentation* is required: The CAPTCHA images need to be first segmented into single alphanumeric characters, thus a more complex image preprocessing is required than previous.

Segmentation-based methods: KNN

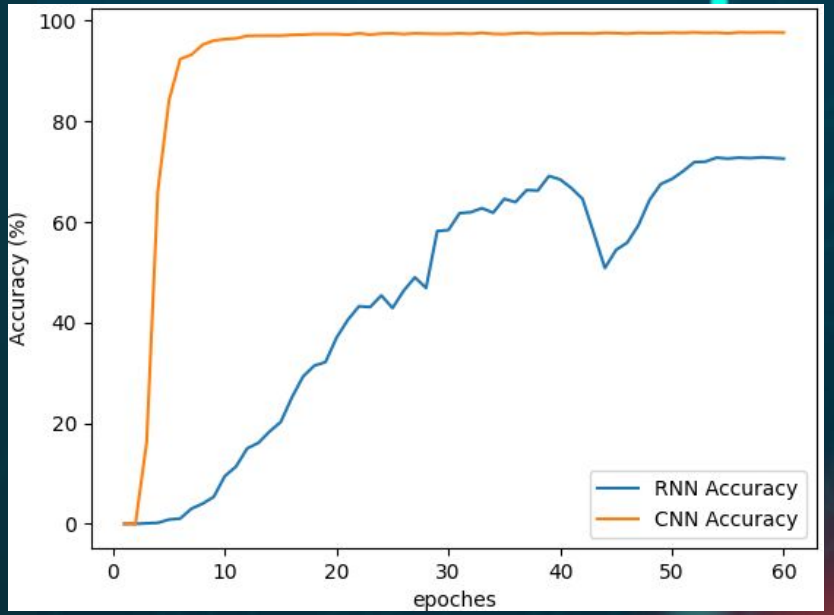
- k-Nearest Neighbours
- Tool used: `sklearn.neighbours.KNeighborsClassifier`
- Classifies each character based on the most common class of its k closest distance neighbours
- Distance is calculated using Euclidean Distance based on the image pixels



Source: https://morioh.com/p/39a8acaae3ce

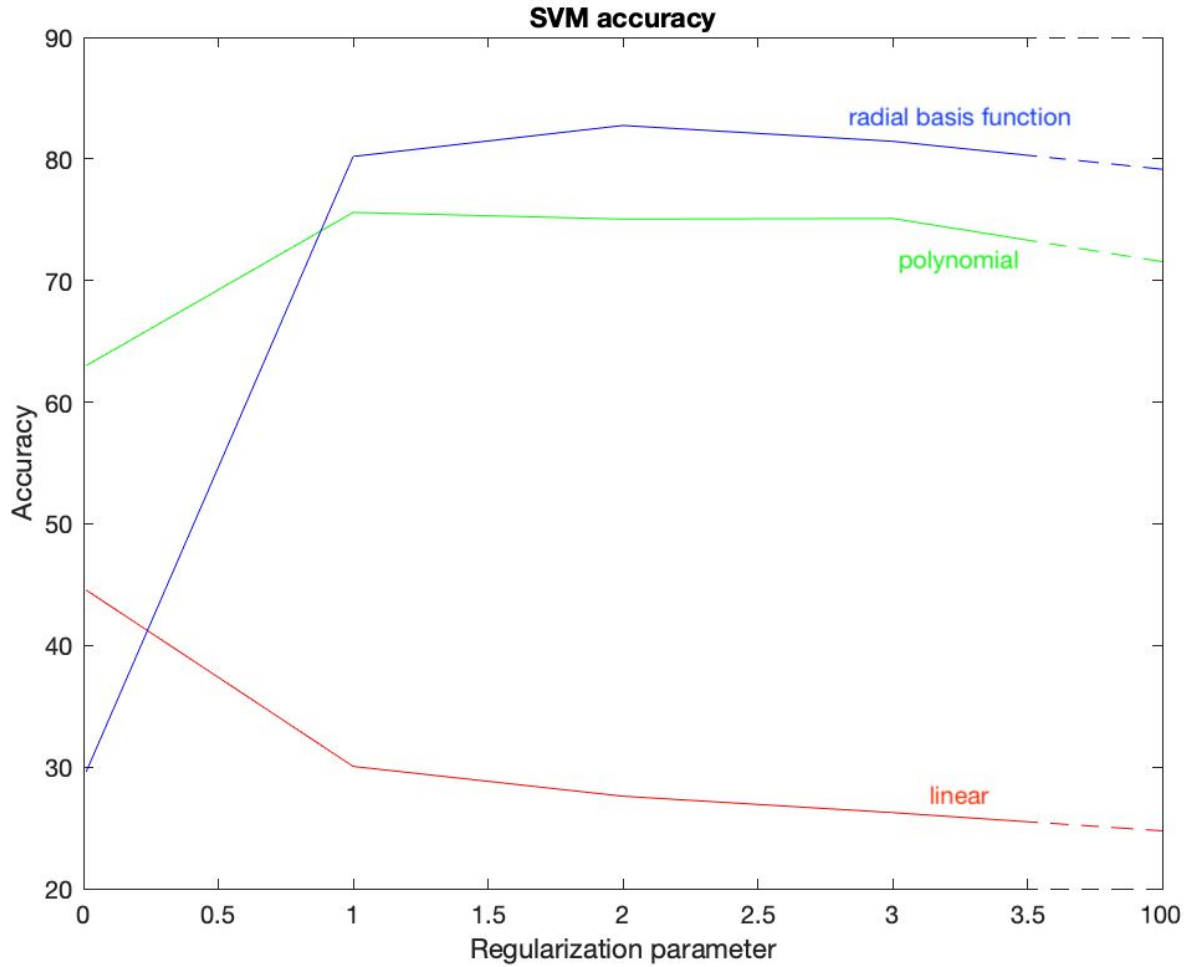
Results: segmentation-free

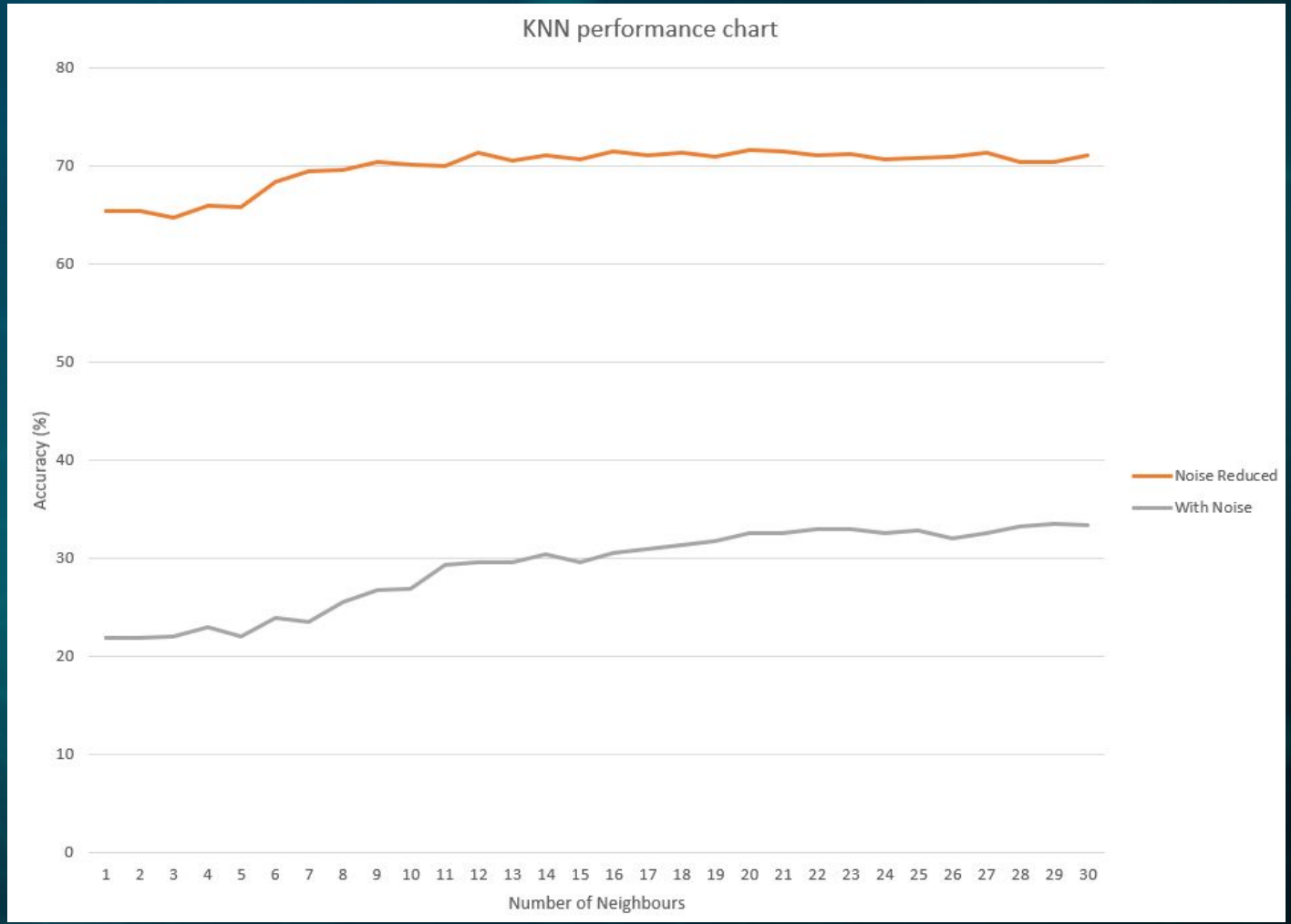
- CNN model has considerably better performance than RNN on this problem
- Possible reason:
  - Spatial properties of CAPTCHA images?



Result: segmentation-based

- For SVM, accuracy can reach around 80%. The *kernel type* plays a particularly important role in the accuracy
- For KNN, maximum accuracy is around 72% with k = 20 neighbours, the algorithm is highly susceptible to *noise*


SVM accuracy
90
80
70
60
50
40
30
20
Accuracy
radial basis function
polynomial
linear
0
0.5
1
1.5
2
2.5
3
3.5
100
Regularization parameter


KNN performance chart
80
70
60
50
40
30
20
10
0
Accuracy (%)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Number of Neighbours
Noise Reduced
With Noise

Discussion, Conclusions, Future work

CNN	RNN	SVM	KNN
Champion Best performance Robust	Hope of variable-length CAPTCHA	Good performance even with rudimentary segmentation algo	Good performance with little noise

Thank You for watching!

Group Members:

- ❑ Yuxi Chen
- ❑ Zijie Tan
- ❑ Lijiangnan Tian
- ❑ Ze Hui Peng

Voice-over: Ze Hui Peng

RESOURCES

1. https://www.kaggle.com/ethan404/captcha6digits
2. https://code.google.com/archive/p/kaptcha/
3. https://github.com/Ethan707/CAPTCHA-Generator