Winter 2021 CMPUT 466 Project Proposal

List of all group members

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

Introduction & Related Work

Introduction/Motivation

Nowadays, CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) has been a popular method to distinguish robots from human users, which has protected our internet security for a long time. From website login to transaction confirmation, CAPTCHA is always around us in our daily life. Typically, CAPTCHAs can be divided into two categories: image-based CAPTCHA, text-based CAPTCHA. In the past few years, CAPTCHA also had a significant change in its form, from the combination of numbers and letters to selecting the right images among several images. But as the CAPTCHA develops, its corresponding recognition technology is also improving. Under this circumstance, we want to focus on the text-based CAPTCHA and implement some recognition models, from traditional algorithms to convolutional neural network models. We will make a comparison and analyze the detailed reasons for the different results. Meanwhile, we also plan to explore the impact of the training data size on the model.

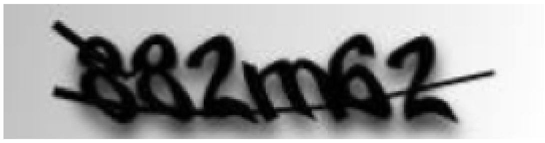
Related work

Traditionally, the text-based CAPTCHA problems can be addressed by the OCR (optical character recognition) technology. For image-text recognition, the traditional method is to first make an object detection for the image, locate the single character, and then divide the image into single characters to recognize. In 2008, Yanping Lv and his team used this method to recognize the Microsoft CAPTCHA and finally achieved 60% accuracy [1]. As machine learning and convolutional neural network fields continue to evolve, better models can be trained to reach higher accuracy. In 2016, Yan and his team used a convolutional neural network model to achieve about 90% accuracy on the Chinese CAPTCHA dataset [2]. And later, the addition of the RNN (Recurrent neural network) and LSTM (Long Short-Term Memory) made the text-based CAPTCHA obtain high accuracy while significantly reducing the model size, which is a huge milestone. In 2017, Baoguang Shi and his team added the recurrent layers after the convolutional neural network and named it the Convolutional Recurrent Neural Network (CRNN). This model just has a size of 8.3M but still have about 97% accuracy on different datasets.[3]

Those researches aim to improve the quality and efficiency of the text recognition models, but they only perform simple analyses of different models and focus on their models. Thus, we want to compare these models to analyze their differences and explain why they make up for the deficiencies of these previous analyses.

Data

Our data is currently comprised of ten thousand CAPTCHA images of size 200 by 50 pixels, each of which has a darkened background and contains a 6-character-long string — a combination of warped digits and deformed letters with some squiggle curves drawn across and alongside. The datasets are either from Kaggle or generated from Google Kaptcha by one of our group members, Ethan Yuxi Chen. Every image file is designated in the format `{content}_{index}.jpg` where `{content}` serves as the corresponding label and contains the actual string in the image which our models need to recognize. In this CAPTCHA recognition problem, which is a supervised classification problem, the labels of our data are multiclass — the number of our labels is the same as the number of images inside our data. The following figure contains one of the CAPTCHA images in the data.



A sample CAPTCHA image with content **`882m62`**

CAPTCHA-6-digits

While considering the accuracy of our models, our models may not predict all of the six characters in an image correctly. Still, the prediction result may be partially correct; some characters are accurately predicted and others wrongly. Therefore, we need to compare the prediction result to the corresponding label character-wisely instead of using the rudimentary string equality comparison.

The interpretability of our models is not crucial since we do not need to know how our models come up with the predictions.

Analysis / Methodology

Most of the currently available text-based CAPTCHA recognition models can be classified into two categories: segmentation-based algorithms and segmentation-free algorithms, according to Thohbani et al.'s research [4]. We are interested in the differences between the effects of these models and algorithms, as well as the reasons behind these differences.

Our plan, at the present stage, is to apply some most commonly-used methods of both categories and their combinations to train text-based CAPTCHA recognition models, compare their performances, and investigate the factors that may affect their performances.

Another potential direction of our project is to work on a specific recognition model focusing on different architectures of that network to compare and study their performances. For example, we can train on networks with architectures CNN + Bi-LSTM + CTC, CNN + LSTM + CTC, and CNN + Bi-LSTM + Attention + CTC, and then investigate their differences and mechanisms [5].

Some algorithms we plan to work on:

1. Segmentation-free algorithms
    a) Convolutional neural network (CNN)
    b) Recurrent neural network (RNN)
2. Segmentation-based algorithms
    a) Support-vector machine (SVM)
    b) *k*-nearest neighbors algorithms (KNN)

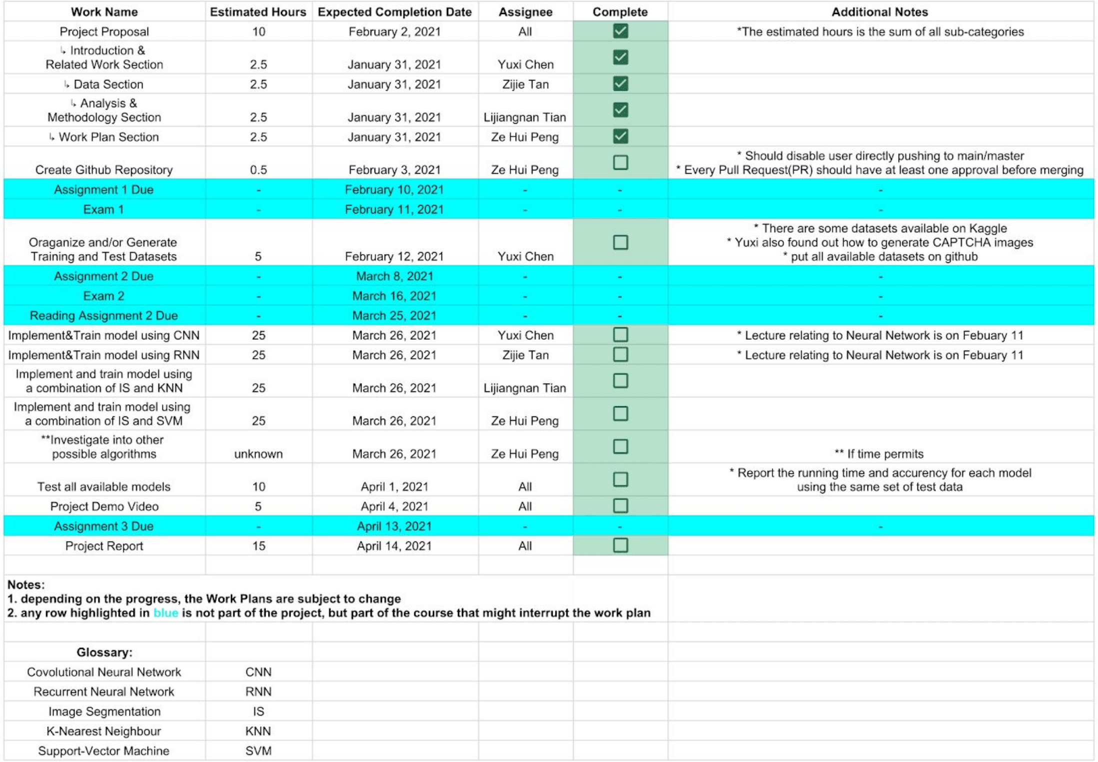
Measuring the performance:

In general, this is a classification problem, and therefore it is natural to use evaluation metrics like classification accuracy/error and F1 score to measure the performances of the models we train. Specifically, for the segmentation-based algorithms, we can introduce extra performance evaluation by calculating their accuracy in recognizing single letters.

At this point, we also believe that techniques like nested cross-validation will be helpful for the evaluation and comparison of models if the dataset is eventually not too large.

Work Plan

Here is a screenshot of our work plan from Google Sheets:



Work Name	Estimated Hours	Expected Completion Date	Assignee	Complete	Additional Notes
Project Proposal	10	February 2, 2021	All	☑	*The estimated hours is the sum of all sub-categories
↳ Introduction & Related Work Section	2.5	January 31, 2021	Yuxi Chen	☑	
↳ Data Section	2.5	January 31, 2021	Zijie Tan	☑	
↳ Analysis & Methodology Section	2.5	January 31, 2021	Lijiangnan Tian	☑	
↳ Work Plan Section	2.5	January 31, 2021	Ze Hui Peng	☑	
Create Github Repository	0.5	February 3, 2021	Ze Hui Peng	□	* Should disable user directly pushing to main/master * Every Pull Request(PR) should have at least one approval before merging
Assignment 1 Due	-	February 10, 2021	-	-	-
Exam 1	-	February 11, 2021	-	-	-
Oraganize and/or Generate Training and Test Datasets	5	February 12, 2021	Yuxi Chen	□	* There are some datasets available on Kaggle * Yuxi also found out how to generate CAPTCHA images * put all available datasets on github
Assignment 2 Due	-	March 8, 2021	-	-	-
Exam 2	-	March 16, 2021	-	-	-
Reading Assignment 2 Due	-	March 25, 2021	-	-	-
Implement&Train model using CNN	25	March 26, 2021	Yuxi Chen	□	* Lecture relating to Neural Network is on Febuary 11
Implement&Train model using RNN	25	March 26, 2021	Zijie Tan	□	* Lecture relating to Neural Network is on Febuary 11
Implement and train model using a combination of IS and KNN	25	March 26, 2021	Lijiangnan Tian	□	
Implement and train model using a combination of IS and SVM	25	March 26, 2021	Ze Hui Peng	□	
**Investigate into other possible algorithms	unknown	March 26, 2021	Ze Hui Peng	□	** If time permits
Test all available models	10	April 1, 2021	All	□	* Report the running time and accurency for each model using the same set of test data
Project Demo Video	5	April 4, 2021	All	□	
Assignment 3 Due	-	April 13, 2021	-	-	-
Project Report	15	April 14, 2021	All	□	

Notes:
1. depending on the progress, the Work Plans are subject to change
2. any row highlighted in blue is not part of the project, but part of the course that might interrupt the work plan

Glossary:	
Covolutional Neural Network	CNN
Recurrent Neural Network	RNN
Image Segmentation	IS
K-Nearest Neighbour	KNN
Support-Vector Machine	SVM

The up-to-date version of our project work plan can be viewed from our Project Work Plan Google Sheet.

References

1. A low-cost attack on a Microsoft captcha. 2008. Accessed February 2, 2021. https://search-ebscohost-com.login.ezproxy.library.ualberta.ca/login.aspx?direct=true&db=edsoai&AN=edsoai.on1098300311&site=eds-live&scope=site

2. Lv Y, Cai F, Lin D, Cao D. Chinese character CAPTCHA recognition based on convolution neural network. 2016 IEEE Congress on Evolutionary Computation (CEC), Evolutionary Computation (CEC), 2016 IEEE Congress on. July 2016:4854-4859. doi:10.1109/CEC.2016.7744412

3. Shi B, Bai X, Yao C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39(11):2298-2304. doi:10.1109/TPAMI.2016.2646371

4. Thobhani A(1), Gao M(1), Hawbani A(2), Ali STM(3), Abdussalam A(4). CAPTCHA recognition using deep learning with attached binary images. Electronics (Switzerland). 9(9):1-19. doi:10.3390/electronics9091522

5. Character-Based Handwritten Text Transcription with Attention Networks. 2017. Accessed February 2, 2021. https://search-ebscohost-com.login.ezproxy.library.ualberta.ca/login.aspx?direct=true&db=edsoai&AN=edsoai.on1106281594&site=eds-live&scope=site