Mina Lee, Lyla Kiratiwudhikul, Tom Zhang
AC209B

Project Proposal

Title:
Implementing an OCR Model and Post-OCR Correction Mechanism from Scratch

Members:
Tom Zhang (szhang1@g.harvard.edu),
Mina Lee (minalee@g.harvard.edu),
Lyla Kiratiwudhikul (lkirati@g.harvard.edu)

Background and Motivation:
Inspired by the classic MNIST handwritten digits classification task, we'd like to expand this idea further by working with a much larger set of data. This is OCR (Optical Character Recognition). Essentially, it extracts text from images (this could be scanned documents or images of handwritings). We would get all kinds of images and train a neural network with them. However, it should be assumed that the OCR model is not perfect; there will be misspellings in its output text, but there are several techniques to correct these mistakes. Our project would then be to train this neural network then implement and compare various post-OCR correction methods, from scratch.

Data:
Honestly, any form of image that contains text would work. We choose the IAM Handwriting dataset[1] to start with. This dataset contains over 1,500 pages of scanned handwritten English text written by 657 writers. Our initial look at the data suggests that the data is relatively clean. The data is labeled at the sentence, line, and word levels, and has been verified manually by the Institute of Complex Systems.

Scope:
As mentioned in the background section, we will implement the model and correction scheme from scratch. If time permits, we could also experiment and compare different configurations for the neural network or use a completely different model. If the training process proves to be more time-consuming than expected, we could restrict the type of data we're feeding in (e.g., only use scanned legal documents). At the end, we would analyze a sample of OCR text output as well as the corrections manually to produce some form of performance metric which could be used to compare different setups.

[1] https://fki.tic.heia-fr.ch/databases/iam-handwriting-database