Optical Character Recognition

Optical Character Recognition, or OCR, is the process by which handwritten characters are parsed into machine-encoded text.

- This is done using a CNN built in Keras.
- Model was trained on EMNIST, a dataset of over 146,000 handwritten letters.
- Model's greatest strength is its 98% top-3 classification accuracy.

Segmentation

Segmentation is the step where lines, words, and letters are separated from each other for processing.

- Lines are separated by looking for horizontal white space.
- Words are separated by increasing the thickness of lines and seeing what is touching.
- Letters are separated by looking for “the most” vertical whitespace.

Word Prediction

Word Prediction pulls information from a variety of sources to piece together the most likely words and put them in order.

- OCR - The 3 most likely letters for each letter cropped from text.
- Segmentation - The ways in which letter data can be used (The graph from the segmentation box).
- A word corpus - A list of words and how often they were used in the data scraped from Wikipedia.

Sentence Prediction

Sentence prediction uses a large corpus of text to predict the likelihood that a word comes before or after another word.

- Word context is determined by leading and following words.
- Module loops through sentence, determining likelihood of each word.
- If likelihood too low, word predictor module is called to generate new possibilities.

Next Steps

If more work was to be done on this project in the future, there are a couple things to keep working on.

- App development - The original project idea was to get this working on a smartphone app.
- More segmentation - A lot of progress has been made this year, but we know that more time will lead to much better results.
- There have been ideas thrown around to add a second neural network to help with this problem.