Logographic Language Detection and Recognition Using a Deep Learning Approach

Yi, Nicholas (School: The Harker School)

Of the estimated 7,100 languages spoken today, half are in danger of extinction by the end of the century, with one language dying out every two weeks. The biggest problem in recording written transcripts of lesser-known languages is that there are a low number of experts for any endangered language. One particular example is with old Japanese cursive text, Kuzushiji, of which over 3 million manuscripts stored and recorded today. This project will use a deep-learning based approach to precisely detect the position and size of each character in Kuzushiji books, correctly recognize each detected character, and properly present them as modern Japanese characters, therefore democratizing ancient manuscripts. The main contribution of this research is creating a context-based image classification neural network model to recognize input characters. After using 80% of the dataset images to train, this model can achieve 97.83% accuracy in recognizing the remaining 20% of character images. As a comparison, a popular and also powerful Xception-based NN model has been evaluated by only using character images without a context feature. Its accuracy is 96.46% with the same dataset. As for the overall system performance, which include both the character detection and character recognition, the F1 score is 0.896. The main challenge of this two-stage approach is that the first stage center-net based character detection is not carefully fine-tuned, it misses to detect some characters, or detects the characters with shifted/reduced/enlarged boundaries, or falsely generates boundaries for the areas without any characters. With the novel context-aware character recognition model in the second stage, the setbacks has been overcome and overall performance is very promising.