Arabic Medieval Manuscript OCR With Seq2Seq

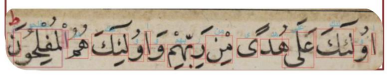
Ammar Alqatri (ammarq@stanford.edu), Nawaf Alnaji (alnajina@stanford.edu)
CS230: Deep Learning

Problem

Optical character recognition on handwritten character remains an open problem, especially in domains such as medieval manuscripts where scripts and page structure are highly variable. Arabic poses a greater difficulty because it's a purely cursive script. We focus on a specific part of the problem. In particular, given an image of an Arabic medieval manuscript sequence, we would like to output its transcript with reasonable accuracy.

Dataset

We use the VML-HD dataset [1]. It is composed of 5 books handwritten between the 11th and 15th centuries, with a total of 680 pages. Each page is annotated at the subword level, with the transcription of the subword and the location of it in the page image. Since Arabic is a cursive script, a subword here refers to a connected component of a word. In total, there are approximately 150,000 subwords, and 12,000 lines.

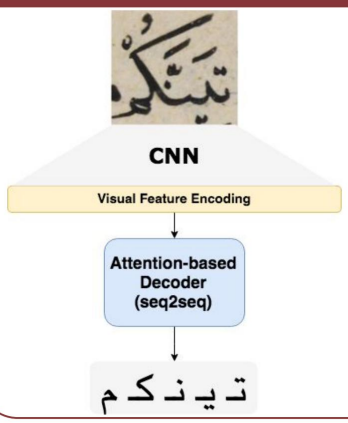


Line Segmentation

We perform line segmentation through statistical image processing rather than deep learning. Our algorithm is a simplified implementation of [2]:

- Binarize image using Otsu threshold
- Find mean number of white pixels per row in the page
- Cluster rows with white above mean
- Set line dividers to cluster centers

Model Architecture



Model Architecture

The model is composed of a CNN component which flattens the input image (which is of fixed height, but variable width) into a 1-dimensional sequence of features. This encoding is then passed to a sequence to sequence model (with LSTM units) with attention which is trained with cross-entropy softmax loss.
We use beam search and a character-level language model we trained on the Quran to obtain the output from the decoder.
We can input both sub-word images and entire lines into the model.

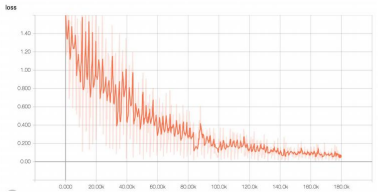
Experiments & Results

Test-set accuracy calculated based on:

- Exact sequence match: %59
- LD (Levenshtein Distance): %71
- LD, disregarding dots: %81

بـتـثـيـنـ

- LD, disregarding dots, with GRU: %74
- LD on entire lines (prelimenary): %13



Future Steps

- Optimizing transcription model for entire line inputs
- Generating a transcript segmented into words. Currently, our dataset does not indicate spaces between words.
- Composing our line segmentation and sequence recognition models to result in an full system which takes pages as inputs and outputs transcripts.

References

[1] Majeed Kassis, et al, "VML-HD: The historical Arabic documents dataset for recognition systems", *(ASAR 2017)*
[2] O. Surinka, et al, "A* Path Planning for Line Segmentation of Handwritten Documents", (ICFHR 2014)