Progress Report

07/31/2019
Mike Pack

Background

Based on the discussion in the kick-off meeting, there are two main tasks I am currently working on:

1. Page segmentation: Aims to identify image-like components—such as cartoons, illustrations, photographs, and maps—from Chronicling America corpus.
    a. *dhSegment* is known to be the state-of-the-art page segmentation algorithm in literature (https://arxiv.org/abs/1804.10371)
    b. The concept of this model is to combine two deep learning models—*ResNet-50* and *U-net*—which are known to be the best model for image classification and pixelwise-classification problem.
    c. Open-source code is provided by the author in GitHub (https://github.com/dhlab-epfl/dhSegment)
2. Metadata generation: Explore various approaches to find what will be the best way to build a well-structured metadata for image-like components.

Discussion of achievements

1. Page segmentation
   1.1. As a pilot experiment, we were able to train *dhSegment* using a small subset (30 images) of **European Historical Newspaper Dataset** (ENP); and obtained a **promising result** (Please see Figure 1).
      1.1.1. We have also confirmed that the model trained on ENP dataset is also capable of separating images from **Chronicling America corpus** into background, text, and figure sub-regions (Please see Figure 2).
   1.2. We have explored the **Beyond Words** JSON data to analyze and construct a well-formed ground-truth set for training the model.
      1.2.1. A script is implemented. This script converts a single JSON file into a number of XML files equal to the number of actual newspaper pages presented in the JSON file. Note that we are using XML format the **PAGE XML** format is known to be a standard format for newspaper segmentation competition (https://www.primaresearch.org/tools/PAGELibraries).
   1.3. We have confirmed that the training result of the model trained on the Beyond Words dataset is **not promising enough** compared to the model trained on ENP dataset.
      1.3.1. More detailed discussion about this result is described in the following "Discussion of problems" section.

2. Metadata-generation
   2.1. I have not started on this task yet, however, I have shared one idea in our shared folder (https://docs.google.com/document/d/1H0oIUh76_QXslCs_PPvf0lV56zJUot3tza0LjfKdG9U/edit).

Discussion of problems

There are three following main concerns in Beyond Words ground-truth dataset that might cause a model to be hindrance during the training:

1. Inconsistency
   1.1. Not all figure entities presented in a page are annotated. More detailed information is described in Figure 3.
2. Imprecision
   2.1. Most of the time, a simple rectangle annotation contains regions that are not relevant to the corresponding class. More detailed information is described in Figure 4.
3. Data imbalance
   3.1. In the JSON file, the class of most of the figure entities is “Photograph.” With the imbalanced dataset, a model can be biased to learn a set of features relevant to the majority class during the training. More detailed information is described in Figure 5.

Discussion of work that lies ahead

1. Segmentation
   1.1. Training model with Beyond Words dataset to address data imbalance problem
   1.2. Training model with enlarged ENP dataset
2. Meta-data generation
   2.1. Explore techniques to generate meta-data relevant to image quality
   2.2. Explore techniques to generate meta-data relevant to image context

Figures

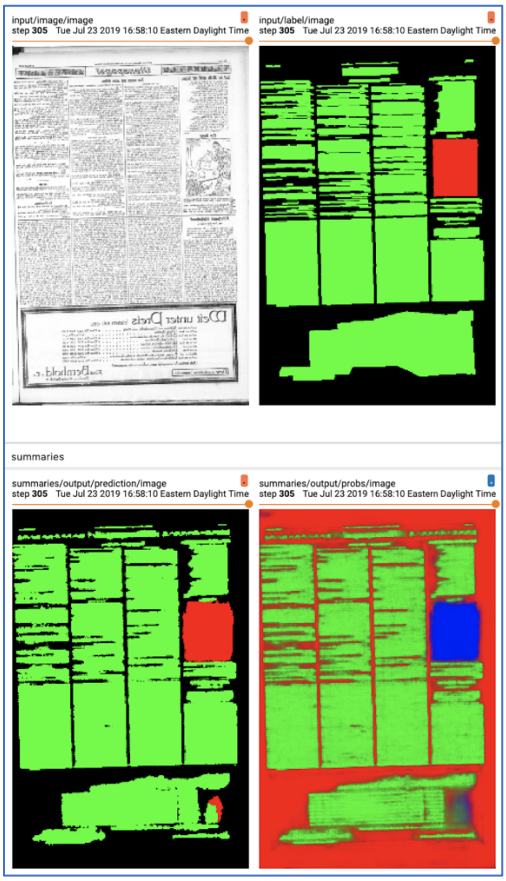

Figure 1. Visual inspection on the segmentation result of model trained on ENP dataset. Clockwise from top-left: (1) Input, (2) ground-truth, (3) probability map, and (4) prediction. In ground-truth, each pixel is labeled as following: (1) black=background, (2) green=text, and (3) red=figure. The probability map here shows the model's pixel-wise prediction value, for example each pixel will have a list of probability values, such as [background:0.2, text:0.7, figure:0.1]. The prediction map is a thresholded result from the probability map, using the arguments of the maxima (i.e., argmax), for example, argmax[background:0.2, text:0.7, figure:0.1]= text:0.7. The color representation of the probability map is the same as the ground-truth.

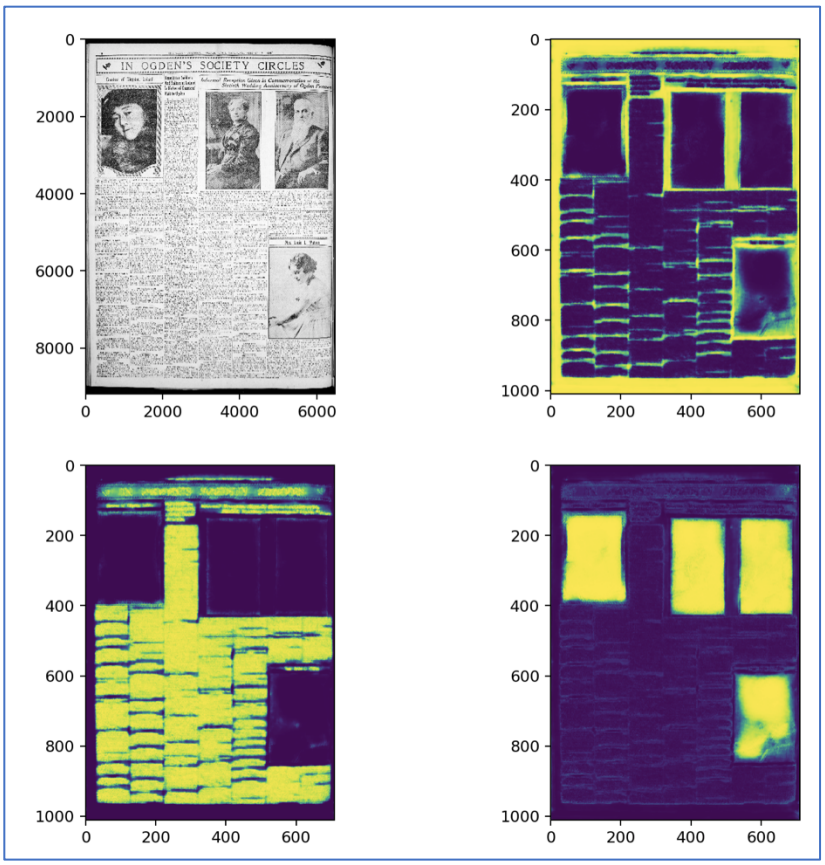

Figure 2. Visual inspection on the segmentation result of model trained on ENP dataset. Note the image shown here is from the Chronicling America corpus, which is never shown to the model during the training. Clockwise from top-left: (1) Input, (2) background-map, (3) image-map, and (4) text-map. In each map, brighter (yellow-ish) region indicates the region of interest with high probability.

Figure 3. Visual inspection on the segmentation result of model trained on Beyond Words dataset. Clockwise from top-left: (1) Input, (2) ground-truth, and (3) prediction. Note here that model makes a reasonable guess that there are multiple figure-like regions in a given page, but the inaccurate ground-truth missing some figure-like regions penalize the model's prediction, which is problematic since it will confuse the model to learn a set of useful features.

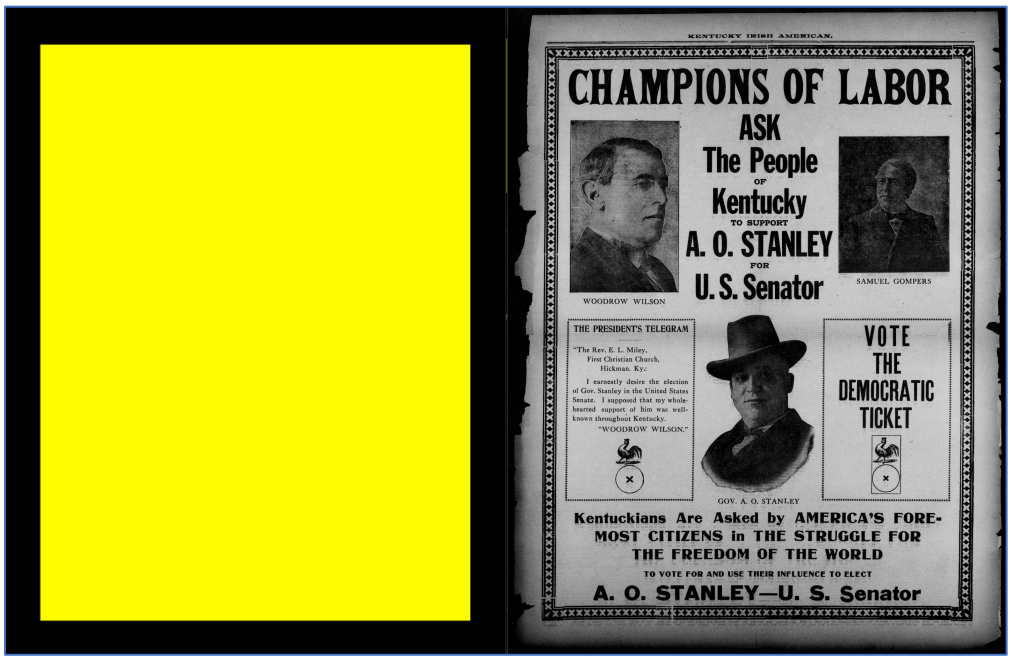

Figure 4. From left to right: (1) ground-truth (yellow: Photograph and black: background) and (2) original image. Note here that in the ground-truth, non-photograph-like (e.g., texts) components are included within the yellow rectangle region. The best-case scenario is to have a more accurate annotation with polygon so that each ground-truth entity can contain only photograph-like pixels.

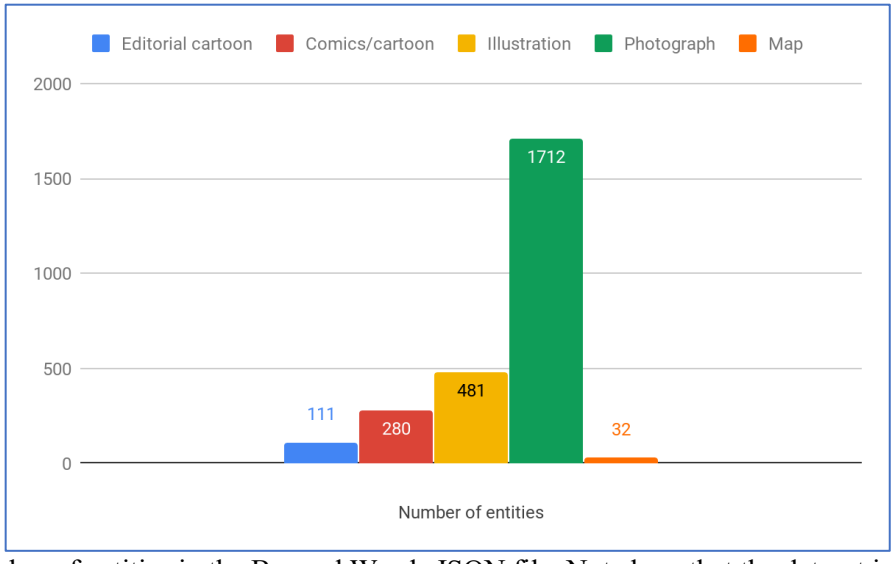

Figure 5. Number of entities in the Beyond Words JSON file. Note here that the dataset is overwhelmed with photograph class (4%, 11%, 18%, 65%, and 1%).