Document Classification

By: Patrick Brown

Case

Student loan servicers have to process thousands of standardized Department of Education forms. Borrowers submit these forms through various mediums. Often, these submissions have to be manually viewed and code for routing based on the form. Can we automate this document coding process?

Design

The goal of this project is to create a classification model to classify document images. There are 3 stages we moved through while developing a solution.

1. Data. There is no publicly available dataset. We created our own by taking template images and using programtic manipulations.
2. Image processing. A raw image is extremely large, noisey and overall ineffecient for processing. We converted images to grayscale before applying thresholding to binarize the pixel data. Last we resize the image to a smaller standard (200x200).
3. Modeling. We explored numerous modeling solutions for this problem. Ultimately, we found that a Bernoulli Naive Bayes model fed raw binarized image data proved most effective.

Data

Ultimately, our dataset was dynamically generated. Training cosisted of 204,000 generated images. Our process involved taking template forms using OpenCV to add random text. Images were placed on random background images with randomized offsets. Lastly, images were put through random augmentations, including but not limited to bluring, masking, and noising.

Tools

- OpenCV
- Numpy
- Albumentations
- Sklearn

Conclusions

Bernoulli Naive Bayes proved surprisingly effective for classifying document images. We found that using an ensemble approach worked best. This approach allowed that model to recognize and classify images as unknown or needing to be reviewed.

Communication

End results were presented via a slide presentation. The champion model developed has been deployed via a Gradio App hosted on HuggingFace, here.