Approach Discussion.

I fine-tuned a trOCR model that was originally trained on handwritten texts from the Rodrigo dataset. The Rodrigo dataset comprises approximately 20K lines and 231K running words sourced from a lexicon of 17K words found in a Spanish manuscript dating back to 1545. This dataset bore some resemblance to the specific dataset provided, hence my choice to utilize it.

After fine-tuning on these text lines, I achieved a Character Error Rate (CER) of 5.4% after 4 epochs, which can be further reduced if necessary, although it is satisfactory for current purposes. The training Jupyter notebook file is provided in the repository under the name "gsoc-ocr.ipynb".

However, the primary challenge I am encountering is the segmentation of text lines from the manuscript pages. Despite attempting various techniques, I have been unable to achieve satisfactory segmentation thus far. I have included one of the segmentation techniques in my GitHub repository for reference.

Following the successful segmentation of text lines on the pages, the lines will be inputted into the trOCR model, which will provide the transcription for each line. Currently, I have deferred fine-tuning the model on the specific dataset provided, as I was unable to segment the lines properly.

Segmentation can potentially be addressed by employing YOLO or Fast R-CNN techniques. However, this approach would necessitate considerable time investment in creating bounding boxes for training these models on the manuscript pages to detect the text lines.

In summary, while the transcription aspect of the project has been completed, the segmentation part remains unresolved and requires further attention. This can potentially be addressed through the outlined approach.