SSY340 - Project Planning

İpek Korkmaz, Yusheng Yang, Jacob Bredin, Muhammad Uzair

October 9, 2024

1 Introduction

The aim of the project is to investigate and create a model that can perform text localization and extraction from images. This task has many applications in the real world and functions as a catalyst for many other technologies and applications.

A notable use-case is in translation software that can translate text in real-time by simply pointing your phone at it. Tools can be made that can be of aid to those with a vision impairment or to people with disabilities that makes it difficult to enter text into a computer. Many tedious tasks that could only be done by humans before, such as data entry or text transcriptions, can now be automated. This can help preserve documents through digitization that would otherwise be lost. This is only a small set of things the technology can be used for.

2 Data Set

For this project, we plan to use a dataset from Kaggle called *TextOCR - Text Extraction from Images Dataset*. The dataset contains a diverse set of images where each text element has been annotated with a bounding box to fit the texts shape. Figure 1 shows some examples of these annotations.

The dataset contains 21k images with approximately 1 million annotations. These images come from various contexts and include texts in different shapes, such as coins, handwritten notes, and posters.

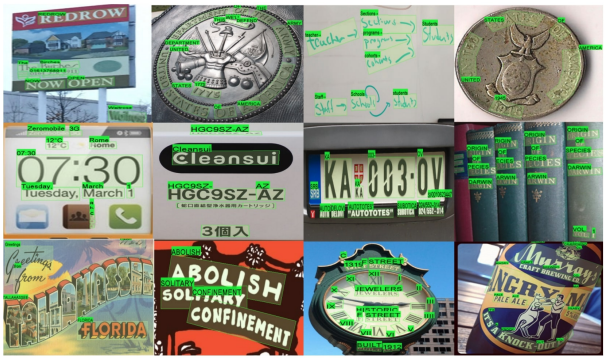

Figure 1: Bounding boxes of text in the TextOCR dataset.

3 Available Code

There are several existing libraries for text recognition, such as *Tesseract* [2] and *EasyOCR* [3], which can handle the recognition task efficiently. We plan to use one of these libraries for the recognition stage. However, the detection part will involve building and training a custom neural network model. The website *paperswithcode.com* has a list of papers that has used the same or similar datasets to ours and we plan to refer to the code of these papers as a basis for our detection model.

4 Relevant Papers

For this machine learning task we have looked at papers that describe methods for object detection and localization, as well as text sequence generation and OCR.

- *You Only Look Once: Unified, Real-Time Object Detection*: Introducing YOLO (You Only Look Once), a new approach to object detection that uses a convolutional neural network (CNN) architecture inspired by GoogLeNet [5].
- *EAST: An Efficient and Accurate Scene Text Detector*: Introducing text detection algorithm called EAST that uses neural network model for predicting words in images [6].
- *Towards Unified Scene Text Spotting based on Sequence Generation*: Presenting UNITS (Unified Scene Text Spotter), a model for text spotting that includes quadrilaterals and polygons. It uses a multi-way transformer decoder alongside a Swin Transformer as the image encoder [7].
- *PIX2SEQ: A language modeling framework for object detection*: Pix2Seq is a generic object detection framework that uses an encoder-decoder model [8].
- *Real-time Scene Text Detection with Differentiable Binarization* Introducing a module called Differentiable Binarization (DB) within CNN for detecting arbitrary-shaped scene text [9].

These papers will guide the design and optimization of the model.

5 Evaluation of Results

To evaluate the model, we will use a test dataset and apply metrics to assess its performance for bounding boxes and text extraction. These will include L2 loss for bounding boxes, and Levenshtein Distance, Character Error Rate (CER) and Word Error Rate (WER) for measuring how well text extraction performs. Additionally, we will use precision, recall, and F1-score for the text detection.

6 Time Plan

- Study week 6. (Thursday, Friday) Review literature, explore the dataset and select a model.
- Study week 7. Implement, iterate and evaluate the selected model.
- Study week 8. Write project report and create poster.

References

[1] R. Mulla. (2022). *TextOCR - Text Extraction from Images Dataset*. kaggle.com. https://www.kaggle.com/datasets/robikscube/textocr-text-extraction-from-images-dataset

[2] Google Inc. (2024). *Tesseract OCR*. Github. https://github.com/tesseract-ocr/tesseract

[3] Jaided AI. (2024). *EasyOCR*. https://github.com/JaidedAI/EasyOCR

[4] A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba and T. Hassner.(2021). *TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text*, The Conference on Computer Vision and Pattern Recognition.

[5] J. Redmon, S. Divvala, R. Girshick and A. Farhadi. (2016). *You Only Look Once: Unified, Real-Time Object Detection*. arXiv.org. https://arxiv.org/abs/1506.02640

[6] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017, July 10). *East: An efficient and accurate scene text detector*. arXiv.org. https://arxiv.org/abs/1704.03155v2.

[7] T. Kil, S. Kim, S. Seo, Y. Kim and D. Kim (2023). *Towards Unified Scene Text Spotting based on Sequence Generation*. arXiv.org. https://arxiv.org/abs/2304.03435

[8] T. Chen, S. Saxena, L. Li, D. J. Fleet and G. Hinton. (2022). *Pix2seq: A Language Modeling Framework for Object Detection.* arXiv.org. https://arxiv.org/abs/2109.10852

[9] M. Liao, Z. Wan, C. Yao, K. Chen and X. Bai. (2019). *Real-time Scene Text Detection with Differentiable Binarization* arXiv.org. https://arxiv.org/abs/1911.08947