Enhancement of OCR service provided by Google Cloud Vision API

Noha Nasser Mohammed Imam	2400060
Mohammed Mahmoud Abd Elftah Elhmadany	2301679

Introduction

Optical Character Recognition (OCR) is not only essential for digitizing printed content but also stands as a practical application of image processing, significantly enhancing productivity and process automation for businesses. Google Cloud Vision API, a leading cloud-based OCR tool, is widely used for extracting text from diverse sources [1]. However, it faces challenges when dealing with complex document structures.

Research Problem and Objectives

The primary research problem is the limitation of Google Cloud Vision API in handling complex OCR scenarios. Specifically, the current OCR functionality struggles with accurately extracting text from documents with multi-column layouts, pages containing intricate tables, and distinguishing between text enclosed within boxes and text outside them. These challenges significantly impact the quality of the extracted text, which is crucial when preparing datasets for training deep learning models and large language models (LLMs). Therefore, the aim of this project is to enhance the Google Cloud Vision API's performance in these complex situations by applying image processing techniques.

Literature Review

State-of-the-art OCR methods utilize deep learning techniques [2], including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [3], to manage variations in text layout, orientation, and complexity. Researchers have developed advanced algorithms like EAST (Efficient and Accurate Scene Text Detector) [4] for detecting text in complex scenes and have utilized Tesseract OCR's adaptive thresholding to tackle challenging backgrounds [5]. In recent years, Transformer-based models have become prominent in OCR tasks, improving the recognition of contextual relationships within structured documents. However, a significant research gap remains in effectively integrating these sophisticated models with cloud-based solutions like Google Cloud Vision API, particularly for precise extraction from intricate layouts.

Methodology

For multi-column texts, we plan to detect the column structure using advanced image processing techniques and apply a separation function for accurate extraction. In the case of tables, we will first detect grid structures to ensure precise extraction of tabular data.

For extracting text that is enclosed within a box, we need to distinguish it from text located outside the box, ensuring accurate differentiation.

These preprocessing enhancements will improve the quality of OCR inputs, boosting extraction accuracy without relying on more expensive alternatives like large language models (LLMs).

Data Collection and Preprocessing

In this phase, we plan to gather a diverse set of images that cover all the scenarios outlined in our proposal. This includes images with multi-column layouts, complex tables, sloped text angles, and text enclosed within boxes. The goal is to comprehensively validate our approach by assessing the quality of text output from these images after applying our preprocessing techniques.

Initially we may use data samples of our own creation to help us build the initial prototype. Once the initial prototype is verified to be fully functional, we may use random datasets from across the internet.

Experimental Design

The main benchmark for this design is to be able to reproduce the digitized documents to be as similar as possible to the original captured document. Ideally our design would enhance the ability of Google Cloud Vision API in reproducing a digital replica of the captured document. Benchmarking the entire project will be performed by determining which areas the system works properly and in which scenarios it suffers.

Expected Results and Analysis

To quantify the effectiveness of the project, we will run the dataset into Google Cloud Vision API twice. Once without any enhancements (control), and another trial preprocessing the documents using image processing techniques (experimental). We expect to improve the API's ability to reproduce digital replicas of the original documents.

Timeline

The table below is a preliminary outline of the timeline of the project. Details of each stage shall be provided in the midterm and Final reports.

Date	Duration	Task
27/10/2024 – 03/11/2024	1 Week	Literature Survey
03/11/2024 – 03/11/2024	1 Day	Dataset Collection
04/11/2024 – 17/11/2024	13 Days	System Architecture Specification First working prototype
17/11/2024 – 28/11/2024	13 Days	Prototype refinement
29/11/2024 – 29/11/2024	1 Day	Midterm report submission
29/11/2024 – 03/12/2024	1 Week	Prepare presentation Final report submission

Potential Contributions

This work can contribute to improving the effectiveness of Google Cloud Vision API, as well as other similar services. Implications of the success of such project means the ability to digitize complex documents layouts, which not only helps in business and consumer productivity, but it also provides access to more data for data-dependent areas such as data science and LLMs.

References

[1] "Google Cloud Vision API Features List," [Online]. Available: https://cloud.google.com/vision/docs/features-list. [Accessed 2024].

[2] N. Subramani, A. Matton, M. Greaves and A. Lam, "A Survey of Deep Learning Approaches for OCR and Document Understanding," 27 November 2020. [Online]. Available: https://arxiv.org/abs/2011.13534v2. [Accessed 25 October 2024].

[3] V. P and F. GM, "Improvement in OCR technologies in postal industry using CNN-RNN architecture: Literature review," *International journal of machine learning and computing,* vol. 12, no. 5, 2022.

[4] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He and J. Liang, "EAST: An Efficient and Accurate Scene Text Detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] T. Hegghammer, "OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment," *Journal of Computational Social Science,* vol. 5, no. 1, pp. 861-882, 2022.