Title: LLM Integration for PDF Generation at CGI

CS4991 Capstone Report, 2024

Aditya Kumar
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
fsc4md@virginia.edu

ABSTRACT
The goal at CGI, a technical consulting firm based out of Canada, is to create a system that automatically converts paper forms into user-friendly digital ones. To accomplish this task, I fine-tuned a large language model, LLaMA, to extract certain feature sets from inputted paper PDFS in order to retrieve all the necessary fields and information needed to begin creating the digital forms. This LLM model is combined with Azure OCR to extract the information, which is then integrated into the existing in-house PDF-creation system. The majority of the coding is done in python via various libraries and API calls. The system is a work-in-progress as the PDF creation process is still separate from the feature extraction system. In addition, to integrate the two, I will need to create a separate system that interprets the data fed into it and then relays the information to the PDF-creation system (i.e., style, placement, layout, specific features, etc). After this is done, I will need to fix bugs and conduct extensive testing.

1. INTRODUCTION

In the technological age, digital transformation is crucial for business efficiency and innovation. Organizations worldwide are seeking effective methods to digitize their operations. The conversion of paper-based processes to digital formats is a crucial part of this transformation. Traditional methods of manually entering data from paper forms into digital systems are time-consuming and inefficient. These challenges highlight the need for automated solutions that can streamline the conversion process, enhance accuracy, and save valuable time and resources. The advent of technologies such as Optical Character Recognition large language models has opened new avenues for tackling these issues. By leveraging these technologies, businesses, in this case, government entities, can automate the extraction of information from paper documents and seamlessly integrate this data into their digital ecosystems.

2. RELATED WORKS

The development of the system at CGI can be compared to research focused on document digitization and information extraction. The integration of machine learning in Optical Character Recognition systems has advanced the capabilities of text extraction from scanned documents. DeepLobe's insights on the modern OCR pipeline highlight the role of machine learning in OCR, offering agility in processing a high quantity of data. This advancement allows for not only recognizing text but also deriving meaningful insights from the information (DeepLobe, 2022). The application of deep learning algorithms also extends OCR's utility beyond just text recognition, facilitating the extraction of text from formats such as hand-printed text, checkboxes, and even low-resolution images, enhancing business workflows by eliminating

manual data entry and improving data processing efficiency.

Also, research by Sharma (2023) details the development of a deep learning architecture for text recognition aimed at addressing the limitations of traditional OCR techniques. Sharma's work highlights the importance of using diverse datasets for training and fine-tuning deep neural networks to ensure performance across various data sources. The study demonstrates the potential of deep learning algorithms to significantly improve text recognition accuracy, adaptability, and speed, even under conditions such as noisy inputs, multilingual text, and complex layouts.

3. PROJECT DESIGN

At CGI, a technical consulting firm based out of Canada, the focus is on using these cutting-edge technologies to address the challenge of digital form creation from paper-based documents. The initiative involves the development of a system designed to automatically convert paper forms into user-friendly, digital formats. The process of transforming paper forms into digital ones involves several steps, including the extraction of data from scanned documents and the integration of this data into a digital form that retains the original's intent and functionality.

By fine-tuning a large language model, specifically LLaMA, CGI aims to extract essential feature sets from paper PDFs, facilitating the retrieval of necessary fields and information for digital form creation. This approach, coupled with the use of Azure OCR for information extraction, will allow CGI to bring paper forms into the digital age. This will then be combined with an ML model on the backend that will interpret the data and create the form.

The first step was to test different OCR models to find one that most accurately pulled the information needed and provided proper labeling. After testing tesseract, AWS, Google, and Azure, we found that Azure had the most robust system and correctly pulled the information needed with the proper labels.

The next step was to locate a suitable LLM model that complied with company security policies and worked well to understand and categorize the extracted data. After tying Claude, LLaMA, and GPT, we found that LLaMA performed the best and complied with data policies set by the company. We then used LLaMA to analyze the data fed into it by the OCR model and categorize the information.

The next step is to create a backend that has premade layouts that can take the information passed into it by the LLM and fill out templates that translate into a digital PDF.

4. RESULTS

Central to our expectations is the successful transition of paper forms into their digital counterparts, with a focus on maintaining the original intent and functionality of each document. This transformation is pivotal, as it not only enhances accessibility but also streamlines the data entry process, making it more efficient and error-free.

The selection of Azure OCR as the primary tool for information extraction is expected to result in a high degree of accuracy in identifying and labeling data from scanned documents. Azure OCR's robust system promises to minimize errors that typically arise during the digitization process, such as misinterpretation of characters or formatting issues. This accuracy is crucial for ensuring that the extracted data is reliable and can be seamlessly integrated into digital formats.

Furthermore, the incorporation of LLaMA as the large language model of choice is anticipated to significantly improve the system's ability to understand and categorize the extracted data. By leveraging LLaMA's advanced natural language processing capabilities, we expect to achieve a high level of precision in identifying relevant features and fields within the scanned documents. This precision will facilitate the creation of digital forms that are not only accurate representations of their paper originals but also optimized for user interaction and data processing.

The backend system, designed to utilize predefined layouts for converting categorized data into digital forms, is expected to be a game-changer. We foresee this system being capable of producing digital forms that are not just visually similar to their paper counterparts but are also enhanced with digital functionalities. This includes features such as dropdown menus, checkboxes, and fillable text fields, which are not possible with paper forms. The successful implementation of this system will mark a significant milestone in CGI's initiative, showcasing the potential for technology to revolutionize document management.

5. CONCLUSION

The development of an automated system for converting paper forms into digital formats is a significant step towards digital transformation and enhanced efficiency for organizations like CGI. By leveraging technologies such as Azure OCR and the LLaMA language model, this project aims to streamline the process of digitizing paper-based documents, reducing manual data entry, and minimizing errors. The combination of Azure OCR's information extraction capabilities and LLaMA's natural language processing is expected to result in a highly accurate and efficient system. This system will not only maintain the original intent and functionality of the paper forms but also enhance them with digital features, such as fillable fields and dropdown menus, which are not possible with traditional paper-based documents.

The successful implementation of this project will provide numerous benefits to CGI and its clients. It will significantly reduce the time and resources required for manual data entry, allowing employees to focus on more valuable tasks. Additionally, the increased accuracy of the digitized forms will lead to improved data quality and reduced errors, ultimately resulting in better decision-making and operational efficiency.

Through this project, valuable insights have been gained regarding the integration of technologies like OCR and large language models in real-world applications. The knowledge acquired during the development process will contribute to the continuous improvement of document digitization techniques.

6. FUTURE WORK

To complete and enhance the current project, several steps need to be taken. First, extensive testing should be conducted to identify and resolve any bugs or inconsistencies in the system. This testing should cover a wide range of paper forms with varying layouts, fonts, and data types to ensure the system's robustness and adaptability.

Second, the integration between the feature extraction system and the PDF-creation system needs to be finalized. This will involve developing a separate system that interprets the data fed into it by the LLM and relays the information to the PDF-creation system, including style, placement, layout, and specific features. Seamless integration

between these components is crucial for the system's overall performance and usability.

Once the system is fully integrated and tested, future work could focus on expanding its capabilities and applications. For example, the system could be adapted to handle handwritten forms or documents in multiple languages. Additionally, the incorporation of machine learning algorithms could enable the system to learn from user feedback and continuously improve its accuracy and efficiency over time.

Another potential avenue for future work is the exploration of other use cases beyond government entities. The technology developed in this project could be applied to various industries, such as healthcare, finance, and education, where the digitization of paper-based documents is crucial for streamlining processes and improving data management.

REFERENCES

DeepLobe. (2022). Machine Learning for OCR: Creating a Modern OCR Pipeline. Retrieved from https://deeplobe.ai/machine-learning-for-ocr-creating-a-modern-ocr-pipeline/

Sharma, P. (2023). Advancements in OCR: A Deep Learning Algorithm for Enhanced Text Recognition. *International Journal of Inventive Engineering and Sciences (IJIES)*, DOI: 10.35940/ijies.F4263.0810823.