Training a Tesseract module for GDR typewriter

Research Lab: Self Reflection Report

Muhammad Kashan Khalid (222100695)

In this research lab, we focused on training Tesseract for GDR Typewriter documents using the German language. The project encompassed several key aspects, including the training of the Tesseract module, its evaluation, and a comparative analysis with established OCR applications like ABBYY Finereader and OCR4all. Additionally, we meticulously documented our findings and summarised the existing work in this domain.

Technical Growth:

Throughout this lab experience, I gained invaluable insights into the intricacies of training Tesseract for Typewritten German documents. I encountered numerous challenges, such as fine-tuning language parameters and ensuring compatibility with the specific fonts used in the documents. Understanding and utilising specific arguments like -l (--lang) and psm (--psm) were essential for optimising the training process. Additionally, I learned the significance of converting PDF files into individual page TIFF files as part of the preprocessing stage, as well as the importance of creating accurate ground truth files to train Tesseract effectively. Furthermore, I acquired knowledge about the role of Box and LSTM files in the training process, which proved instrumental in achieving accurate results.

Moreover, I developed proficiency in evaluating the trained Tesseract module using metrics such as WER (Word Error Rate) and CER (Character Error Rate), and subsequently comparing these results with the performance of ABBYY Finereader and OCR4all. This comparative analysis provided valuable insights into the strengths and weaknesses of each OCR solution.

Contribution:

My contributions to the project encompassed a diverse array of tasks aimed at optimising the performance of Tesseract for GDR Typewriter documents in the German language. I assisted in converting PDF files to TIFF format and played a pivotal role in formulating an effective approach to train Tesseract. This involved thorough research to identify relevant methodologies and techniques for training and evaluation. Additionally, I actively participated in creating ground truth files and generating Box and LSTM files, crucial for providing contextual information to enhance recognition accuracy. Furthermore, I meticulously trained Tesseract with specific parameters tailored to our project's requirements, optimising its performance.

Following the training phase, I conducted a comprehensive evaluation of the trained Tesseract model, comparing its performance with other OCR applications like ABBYY Finereader and OCR4all using metrics such as WER and CER. This comparative analysis provided valuable insights into the strengths and weaknesses of our model, guiding further refinement efforts. Moreover, I undertook the

responsibility of proofreading the documentation to ensure clarity and accuracy in communicating our project findings and methodologies effectively.

Teamwork and Collaboration:

The success of our project was greatly facilitated by effective teamwork and collaboration. Each team member actively participated and contributed to the project's progress. Through open communication and a shared commitment to our goals, we were able to overcome challenges collectively and maximise our potential. This collaborative approach fostered a positive and productive environment, ultimately leading to the successful completion of our project objectives.

Conclusion:

I have gained invaluable knowledge throughout this research lab on training the Tesseract module for GDR Typewriter, enriching my understanding and skills in this domain. I wish to extend my sincere appreciation to Prof. Dr. Jens Dörpinghaus for his thorough guidance and support during our journey. His exceptional expertise was pivotal in the success of our project, and I am grateful for his assistance in transforming this research into a fulfilling and enlightening experience.

Sincerely,
Muhammad Kashan Khalid