Integration of Deep Computer Vision Foundation Models for Document Interpretation:

Enhancing Optical Character Recognition Using an OCR-free Transformer Model

Student

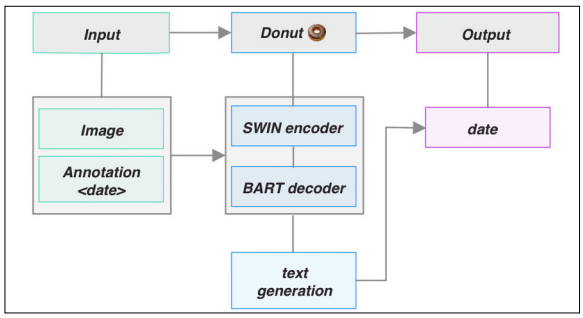
Anastasiia Graftceva

Introduction: This study explores the efficacy of a Hugging Face pre-trained transformer model for Optical Character Recognition (OCR), specifically for date extraction from scanned documents. Initially, OCR relied on pattern recognition and rule-based algorithms to identify text. Deep learning has since enhanced this, enabling handling of complex texts and layouts. This advancement, coupled with Large Language Models (LLMs), has revolutionized visual document understanding.

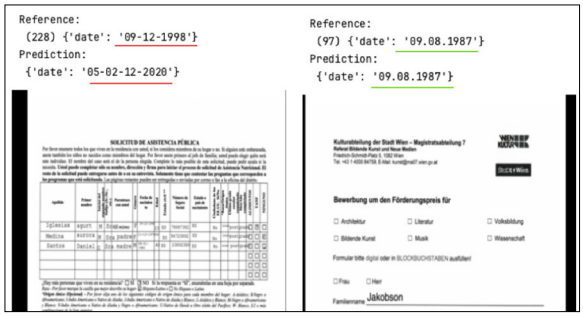
Approach: The conventional OCR approach follows two steps: First, one would OCR a scanned document with the help of an OCR engine like Tesseract, and then process the output using pattern matching and regular expressions, or, alternatively, a LLM trained for the specific field of application. A major limitation of OCR engines, however, lies in their generic nature, which often brings challenges in accuracy and efficiency.
The pseudo-OCR approach instead relies on a single encoder-decoder transformer model which integrates the aforementioned two steps, making it an end-to-end solution which can be adjusted and fine-tuned for a specific field of application.
For my project, I selected the OCR-free Document Understanding Transformer model (Donut), pre-trained on a diverse document dataset. I fine-tuned this on various-sized datasets to assess its proficiency in reading, understanding, and extracting dates from images. I evaluated the model's accuracy and its flexibility with different document types, qualities, and date formats.

Conclusion: The results of the study are encouraging, achieving an average accuracy of 75% on the somewhat limited training and test datasets meticulously assembled for fine-tuning. The OCR-free method shows potential for atomic tasks such as date extraction from images but could benefit from more diverse document types and date formats. Its ability to handle documents with varying numbers of dates per page is a probable area for improvement.

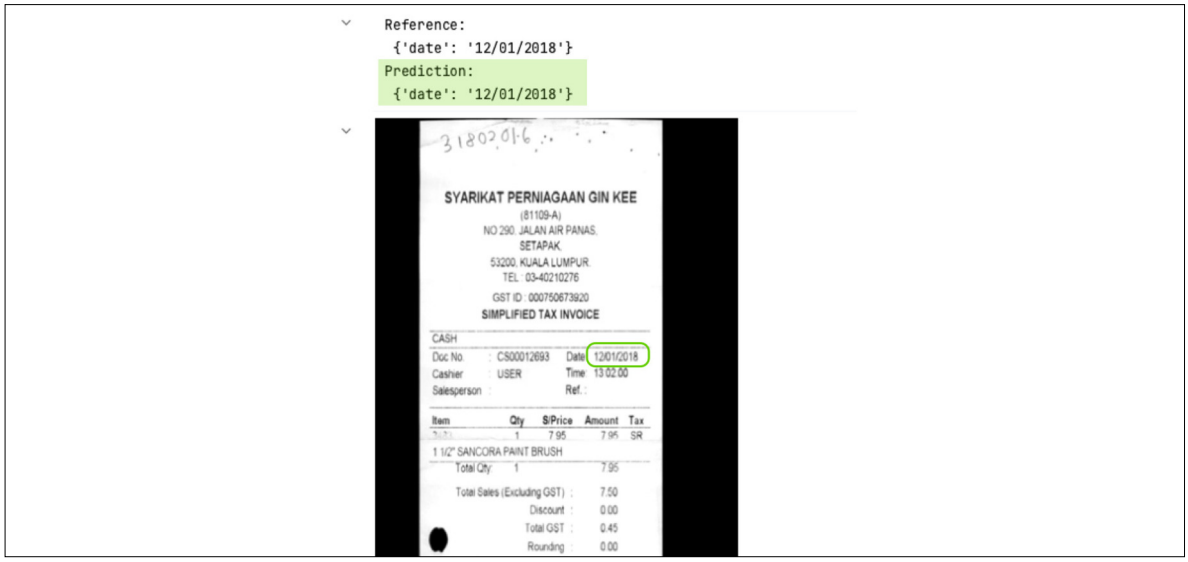
Date extraction using Document Understanding Transformer
Own presentment



Prediction on two types of application form layouts
Own presentment



Correct prediction of the date from an image of the receipt
Own presentment



Advisor
Prof. Dr. Marco Lehmann

Subject Area
Miscellaneous, Software