Document information extraction using layoutLM,Donut

DocVQA (Document Visual Question Answering)

is a research field in computer vision and natural language processing that focuses on developing algorithms to answer questions related to the content of a document, like a scanned document or an image of a text document. Unlike other types of visual question answering, where the focus is on answering questions related to images or videos, DocVQA is focused on understanding and answering questions based on the text and layout of a document. The main challenge in DocVQA is understanding the document's context with layout and formatting to answer the questions accurately.

Benefits Offered by DocVQA

DocVQA offers several benefits compared to OCR (Optical Character Recognition) technology.
Firstly, DocVQA can not only recognize and extract text from a document, but it can also understand the context in which the text appears. This means it can answer questions about the document's content rather than simply provide a digital version.
Secondly, DocVQA can handle documents with complex layouts and structures, like tables and diagrams, which can be challenging for traditional OCR systems.

Challenges Associated with DocVQA

There are several issues and challenges associated with document question answering, including:

- **Understanding the Context:** One of the biggest challenges in document question answering is understanding the context of the document. It is essential to understand the layout, formatting, and language used in the

document to answer the questions accurately. This requires models that can handle the document's structure and content complexity.

- **Ambiguity:** Another significant issue in document question answering is ambiguity. Documents may contain ambiguous or vague language, making it difficult for models to interpret the meaning of the text accurately. This requires models that can handle the nuances of natural language and distinguish between different definitions of the same word or phrase.
- **Limited Training Data:** There need to be large-scale annotated datasets for document question answering, making it challenging to train accurate models. This requires models that can learn from limited amounts of training data and can generalize to new documents.
- **Complex Questions:** Document question answering may involve difficult questions requiring accurate reasoning and inference. For example, a query may require combining information from different document parts to arrive at the answer. This requires models that can perform complex reasoning tasks.
- **Multi-modal Understanding:** Some documents may contain text and images, making it essential for models to have multi-modal understanding capabilities to answer questions accurately.

Related works

DocVQA is a technology developed by several companies and research institutions. The most notable companies working on DocVQA technology include Google, Microsoft, IBM, and Amazon.
Google has document AI, Microsoft has the LayoutLM model, IBM has developed a DocVQA system called the IBM Watson Discovery service, and Amazon has also developed an Amazon Textract service, which can extract text and data from scanned documents, PDFs, and images using machine learning.

LayoutLM

LayoutLM is a pre-trained model for document image understanding developed by Microsoft Research. It is based on the BERT architecture and trained on a large-scale document image dataset to understand document layout, structure, and content.

LayoutLM can be used for various document understanding tasks, including document classification, information extraction, and visual question answering (VQA). In particular, it has shown promising results in the field of DocVQA.

Several research studies have shown that LayoutLM outperforms other state-of-the-art VQA models on benchmark datasets for DocVQA, indicating its potential for practical applications in document understanding.

The Form Understanding in Noisy Scanned Documents (FUNSD) dataset is a benchmark dataset for form understanding and analysis in the domain of noisy scanned documents. It contains 199 real-world scanned document forms and is designed to challenge models in understanding and extracting information from these types of documents.
The FUNSD dataset includes various types of information, like questions, answer and is annotated at both the block and token levels. The annotations are provided in a standard format, including both the ground truth labels and the positions of the blocks and tokens.
The annotations in the FUNSD dataset have been used to train and evaluate state-of-the-art models for form understanding, like LayoutLM and other pre-trained language models. The publicly available dataset is intended to serve as a benchmark for researchers and practitioners interested in form understanding and analysis in the domain of noisy scanned documents.

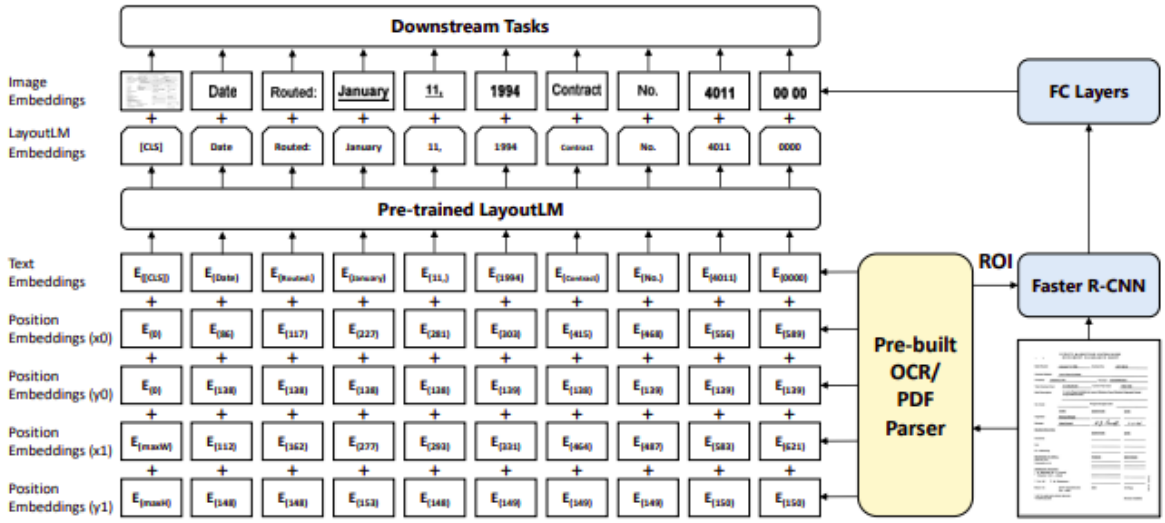

Figure 2: An example of LayoutLM, where 2-D layout and image embeddings are integrated into the original BERT architecture. The LayoutLM embeddings and image embeddings from Faster R-CNN work together for downstream tasks.

DONUT

The DONUT model is proposed in the OCR-free Document Understanding Transformer(DONUT) category by Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, Seunghyun Park. Donut includes an image Transformer encoder and an autoregressive text Transformer decoder to understand the document. It is widely used for Image classification, form understanding, and visual question answering.

DONUT proposes to solve the following problems.

- CORD (form understanding)
- DocVQA (visual question answering on documents)
- RVL-DIP (document image classification)

DONUT consists of a Transformer based visual encoder and textual decoder modules. In which visual encode extracts relevant features from document and

textual decoder convert the derived features into a sequence of sub word to generate a desired output. DONUT doesn’t relay on any of the OCR-module.