Optical Character Recognition(OCR) Extraction using LayoutLM for Document Image Understanding

Umesh Kumar Gattem (ugattem@iu.edu)

Project Overview :

Extracting data from Business Documents through manual efforts is time consuming and expensive, meanwhile requiring manual customization or configuration. The exact format of the document may vary, but the information is usually presented in natural language. It can be organized in a variety of ways from plain text, multi-column layouts, and a wide variety of tables, forms or figures and are in a variety of categories, including letter, memo, email, file folder, form, handwritten, invoice, advertisement, budget, news articles, presentation, scientific publication, questionnaire, resume, scientific report, specification, and many others, which is ideal for large-scale self-supervised pre-training

Earlier a lot of models like Convolution Neural Networks were proposed to extract data(like table detection) from PDF documents, later advanced models like Faster R-CNN model and Mask R-CNN were introduced to further improve the accuracy of document layout analysis. In addition, fully convolutional networks were presented for extracting semantic structures from document images, taking advantage of text embeddings from pre-trained NLP models. Later, Graph Convolutional Networks(GCN) based models were proposed to combine textual and visual information from the documents. To this end, LayoutLM models were introduced, inspired by the BERT model where input textual information is represented by text embeddings and position embeddings. LayoutLM further adds two types of input embeddings: (1) a 2-D position embedding that denotes the relative position of a token within a document; (2) an image embedding for scanned token images within a document. LayoutLM is the first model where text and layout are jointly learned in a single framework for document level pre-training.

Goals :

- Understand the LayoutLM model architecture, train the model and check the performance of the different kinds of dataset.
- Understand different models(like CNN, Faster R-CNN, GCN) which are used to extract data and compare the result with the LayoutLM.
- Test the documents with different layouts and formats and also the complex template structures where documents are of poor quality.

Challenges :

- Although BERT-like models become the state-of-the-art techniques on several challenging NLP tasks, they usually leverage text information only for any kind of inputs. When it comes to visually rich documents, there is much more information that can be encoded into the pre-trained model.
- Few models rely on a few human-labeled training samples without fully exploring the possibility of using large-scale unlabeled training samples. They usually leverage either pre-trained CV models or NLP models, but do not consider a joint training of textual and layout information.
- Extracting data from the poor quality of scanned document images or from the complex template structures is a very challenging task due to the diversity of layouts and formats.

Datasets :

- LayoutLM was pre-trained on IIT-CDIP Test Collection 1.0, which contains more than 6 million documents, with more than 11 million scanned document images. Moreover, each document has its corresponding text and metadata stored in XML files.
- LayoutLM evaluated their approach using The FUNSD(Form Understanding in Noisy Scanned Documents) Dataset which includes 199 real, fully annotated, scanned forms with 9,707 semantic entities and 31,485 words.
- I am planning to use some complex structured dataset where it is difficult to extract the data due to the diversity of layouts and formats.
- One such dataset is Wantok Niuspepa Dataset which contains the News papers readings for several Years.
- In addition to this I wanted to extract data from the complex Documents(like Employee Payslip Dataset, Student Resume/Curriculum Vitae Dataset) and create a Tabular form of the data.

References :

- LayoutLM: Pre-training of Text and Layout for Document Image Understanding
- https://wantokniuspepa.com/index.php/archives
- https://www.kaggle.com/datasets/cityofLA/city-payroll-data
- https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset