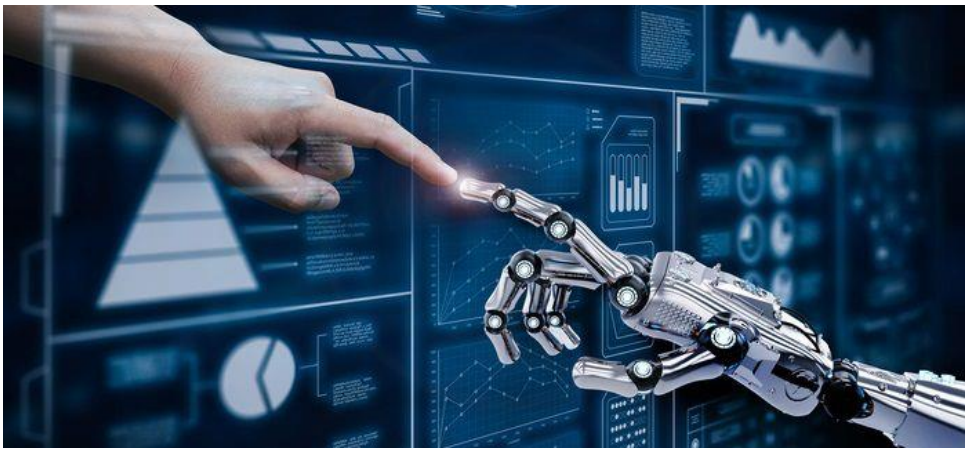

DS ASSIGNMENT REPORT

Prepared by: Saeed Ahmad

Introduction:

First of I would like to thank Nityo Deep Tech for providing me an opportunity to interview and assessment. The assignment was very well designed, although they all are full time projects but I have tried to deliver my best. This assignment has given me opportunity to learn and explore things that I haven't encountered before. It was really hectic though because I have to learn and implement in shortest possible time.

Task 1: **Text Extraction from PDFs.**

I have approached this problem two ways.

1. **Pymupdf library.**
   Pymupdf does pretty decent job on extracting the text, due to shortage of time I was unable to explore the whole functionality of the library. However, I have tried my best to produce good results.
2. **Deep Learning Approach.**
   To effectively achieve the objective of extracting the text in the layout of the document. RCNN based models to draw the bounding boxes around of the text are be.
   Detectron 2 includes several models that are trained on huge dataset to detect Text, Titles and Tables in pdfs. I tried to sort the coordinates of bounding boxes by splitting the page into three columns by setting thresholds on X coordinate values. Then all the boxes in that column were sorted according to their Y coordinate values to extract the document layout. The text was extracted using Google's OCR model from each region.

Task 3: **NER:**

1. **Bert Based Model from Simple Transformers Library:**
   This model uses Google's pretrained BERT encoder for encoding the text. Then the output encodings are classified by a softmax layer. The training only trains the last softmax layer. It produces reasonable results. Around 80% accuracy which is reasonably good. Keeping in mind that the data is poorly annotated.
2. **Spacy Model.**
   In the second model I trained a small spacy model on the data. This model is CNN bases. It also produced reasonable results, to improve we can use bigger transformer based spacy models. Transformers are the best result producing models for NLP tasks.
3. **Further Possibilities.**
   - We can train our own deep learning model. This data was fairly small and simple but for bigger data sets we can use custom built models which use transformers e.g BERT as encoders. The encodings from the BERT or any other pretrained transformer can be passed through 1DCNN, LSTM, Bi-directional

LSTM etc. Usually, the custom deep models in combination with transformers work better on custom NER tasks.

Task 1: **Table Extraction:**

1. Table extraction is one of the hot topics in Computer Vision and Pattern Recognition. A paper was published in CVPR, the biggest CV conference by DevashishPrasad which presented a RCNN and CNN based models to extract table layout. I wasn't able to explore it in great detail but in my opinion, we can use the layout box coordinates to extract structures of the tables. Since bounding boxes around element in same rows have almost similar Y coordinates, we can sort them and extract the structure. Similarly column layout by sorting w.r.t X. Notebook (from Github) has been added.
2. Tabula to extract tables. It extracts data in 3D array. Which can be parsed to extract the layout and structure.
3. We can Use Detectron 2 to draw bounding boxes around tables. Then crop the bounding boxes. These images now can be fed to OCR models to draw bounding boxes on each entry. Then we can use the coordinates of each entry to sort and extract table layout