NLP project's report: Documents layout analysis.

EDDARBALI Niama
AIT BIGANE Youssef

January 10, 2024

Abstract

Document Layout Analysis (DLA) serves as a crucial preprocessing stage in document understanding systems, playing a pivotal role in detecting and annotating the physical structure of documents. The primary goal of DLA is to facilitate subsequent analysis and recognition phases by identifying homogeneous document blocks and establishing their relationships. As of now, a universal DLA algorithm capable of accommodating all document layouts or meeting all analysis objectives remains elusive. In this report, we present an intriguing document layout analysis model called Language-independent Layout Transformer (LiLT)(3) combined with XLM-RoBERTa(5), a multilingual RoBERTa(6) model (trained on 100 languages),both based on the revolutionary Transformers library and finetuned on the famous Doclaynet dataset that contains more than 80000 human-annotated documents from various sources and was conceived for document layout segmentation.

1 Introduction

First of all, as for any Machine learning or Deep learning project, understanding the problematic to solve or the object to achieve are crucial steps to take before embarking the Data understanding, collection, cleaning and processing.

1.1 Problematic

The main goal of this project is to conceive a convenient reliable model for structured documents understanding, regardless of the nature of the document or its language. We also put our lack of high-end hardware resources in mind when doing our research for what's going to serve our purpose.

1.2 Solution

To achieve our goal, we will be needing a really large and reliable dataset that holds diverse categories of structured rich documents with convenient,trustful and preferably humanly annotations, along with a pretrained model that allows the combination of any pre-trained RoBERTa text encoder with a lightweight Layout Transformer, in order to activate a LayoutLM-like document understanding for many languages.

2 Dataset

2.1 Raw Dataset

- DocLayNet (**?**) dataset (IBM) provides page-by-page layout segmentation ground-truth using bounding-boxes for 11 distinct class labels on 80863 unique pages from 6 document categories.
- About the PDF languages: he vast majority of documents contained in DocLayNet (close to **95%** are published in **English** language. However, DocLayNet also contains a number of documents in other languages such as **German (2.5%) French (0.1%) and Japanese (1.0%)**.
- About the categories distributions: The pages in DocLayNet can be grouped into six distinct categories, namely **Financial Reports, Manuals, Scientific Articles, Laws & Regulations, Patents and Government Tenders.** Each document category was sourced from various repositories.

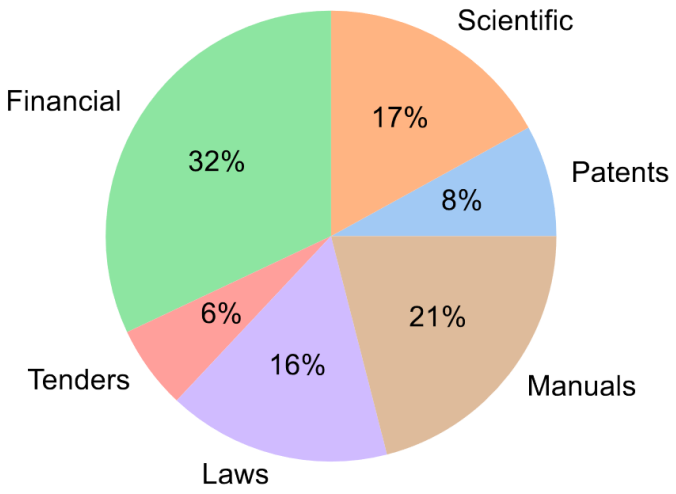


Figure 2: Distribution of DocLayNet pages across document categories.

- Labels of the dataset: The labelling goes the following way:
  - **Text:** Regular paragraphs.
  - **Picture:** A graphic or photograph.
  - Caption: Special text outside a picture or table that introduces this picture or table.
  - **Section-header:** Any kind of heading in the text, except overall document title.
  - **Footnote:** Typically small text at the bottom of a page, with a number or symbol that is referred to in the text above.

- **Formula:** Mathematical equation on its own line.
- **table:** Material arranged in a grid alignment with rows and columns, often with separator lines.
- **List-item:** One element of a list, in a hanging shape, i.e., from the second line onwards the paragraph is indented more than the first line.
- **Page-header:** Repeating elements like page number at the top, outside of the normal text flow.
- **Page-footer:** Repeating elements like page number at the bottom, outside of the normal text flow.
- **Title:** 1 Overall title of a document, (almost) exclusively on the first page and typically appearing in large font.
- **None:** Initial state of each cell/element. Only keep this if the element is not a text or picture or anything else of value. For instance, a smear or an invisible/empty cell should remain "None

2.2 Preprocessed data

- Downloading of all the data (approximately 30GBi) requires downloading time (about 45 mn in Google Colab) and a large space on the hard disk. These could limit experimentation for people with low resources. DocLayNet data had to be processed in 3 datasets accompanied with associated texts and PDFs (base64 format). :

  1. **DocLayNet small** (about 1% of DocLayNet) less than 1000 document images (691 train, 64 val, 49 test)
  2. The DocLayNet base(about 10% of DocLayNet) less than 10000 document images (6910 train, 648 val, 499 test)
  3. **DocLayNet large** (about 100% of DocLayNet) less than 100000 document images (69.103 train, 6.480 val, 4.994 test)

- We opted for the second dataset as the base for our model.

3 Technologies

3.1 Transformers

- The "Transformers" library typically refers to Hugging Face's Transformers library, a popular open-source library for natural language processing (NLP). It is developed and maintained by Hugging Face, a company that focuses on providing state-of-the-art NLP models, datasets, and training pipelines.

- The Transformers library offers a collection of pre-trained models for a wide range of NLP tasks, such as text classification, named entity recognition, machine translation, and more. The models in the library are based on transformer architectures, which have proven to be highly effective in handling sequential data, especially in the context of NLP.

3.2 OCR

- Optical Character Recognition (OCR) is a technology that converts different types of documents, such as scanned paper documents, PDFs, or images captured by a digital camera, into editable and searchable data.

3.3 LiLT

- **Motivation:** with most existing related models restricted and disadvantaged when dealing with the document data of specific language(s) (typically English) included in the pre-training collection, LiLT (3) or Language-independent Layout Transformer came to address the issue. It can be pre-trained on the structured documents of a single language and then directly fine-tuned on other languages with the corresponding off-the-shelf monolingual/multilingual pre-trained textual models.

- Architecture: It be regarded as a parallel dual-stream Transformer. The layout flow shares a similar structure as text flow, except for the reduced hidden size and intermediate size to achieve computational efficiency. The figure above illustrates how LiLT functions:

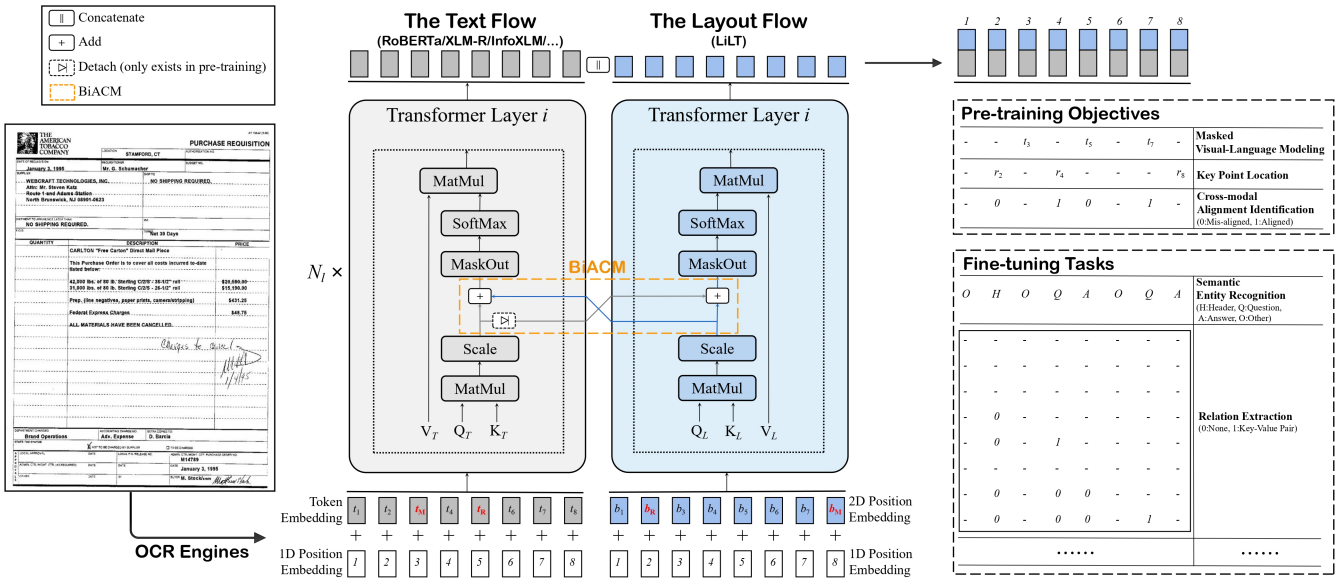


Figure 1: LiLT architecture

It goes by Starting with an input document image, the initial step involves employing off-the-shelf OCR engines to acquire both text bounding boxes and their corresponding contents. Subsequently, the text and layout information undergo separate embedding processes, entering their respective Transformer-based architectures to generate augmented features. To achieve cross-modality interaction between text and layout cues, we introduce the Bi-directional Attention Complementation Mechanism (BiACM)(7) which multi-stage hierarchical process that represents the context at different levels of granularity and uses bidirectional attention flow mechanism to obtain a query-aware context representation without early summarization. In the concluding phase, the encoded text and layout features are concatenated, and ad-

ditional heads are introduced for either self-supervised pre-training or downstream fine-tuning.

3.4 Roberta and XLM-RoBERTa

- RoBERTa (6) is a transformers model pretrained on a large corpus in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts.
- XLM-Roberta (5) is a multilingual version of RoBERTa. It is pre-trained on 2.5TB of filtered CommonCrawl data(vast collection of web pages and content crawled and made accessible for various purposes, such as research, analysis, and development of applications.) containing 100 languages.

4 Model development:

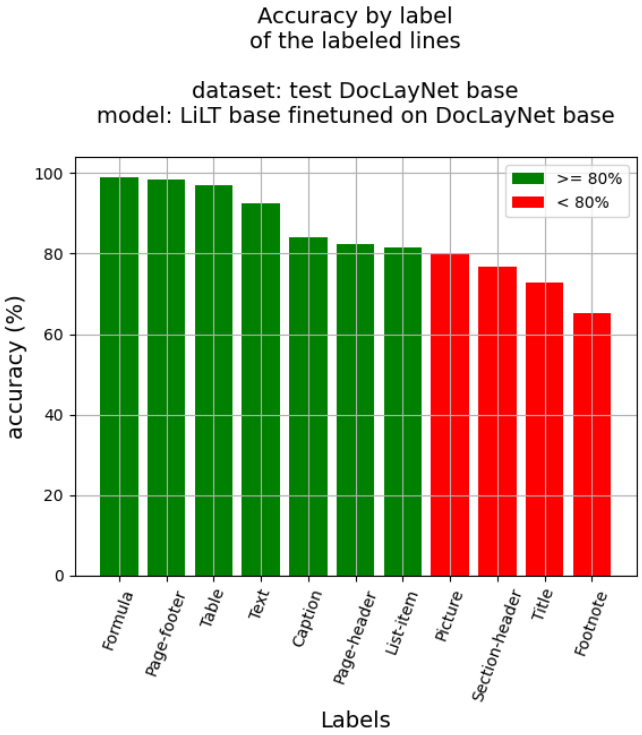
1. We resorted to finetune the following pretrained model which is LiLT + XLM-RoBERTa-base on the DocLayNet-Base dataset.In fact, This model is created by combining the Language-Independent Layout Transformer (LiLT)(3) with XLM-RoBERTa(5), a multilingual RoBERTa(6) model (trained on 100 languages).
2. The finetuning was done on a **line level** with a **chunk of 256 tokens with overlap.**
3. Why chunks? Well, when it comes to **token classification**, dealing with very long documents have us truncate them in other tasks, when they are longer than the model maximum sentence length which very inconvenient in our case since removing part of the the context might result in a worst model.For that, we will allow one (long) example in our dataset to give several input features, each of length shorter than the maximum length of the model (or the one we set as a hyper-parameter). Also, we allow some overlap between the features we generate controlled by the hyper-parameter doc_strid in order to train the model with more contextual information.
4. This data division into chunks with overlap results in a **Encoded dataset**.
5. The encoded dataset is then converted to be compatible with a transformer-based model using the XLM-RoBERTa architecture leverageing the Hugging Face Transformers library for tokenization.
6. This encoded dataset becomes the object of the creation of **PyTorch DataLoaders** in order to ensure that the input sequences are properly padded and organized into batches for training the model.
7. Lastly comes the defintion of the pretrained-model **LiLT + XLM-RoBERTa-base** and so starts the training.
8. The final product which is the finetuned model on DocLayNet-Base dataset, gets pushed to our hub.

5 Inference

- During the inference, We run our own OCR on any PDF document from the test set of our data in order to get the bounding boxes, then load and run our fine-tuned model on the individual tokens and visualize the result at lines level.
- With **Colab** being our workspace, We knew ahead that it might stop the evaluation for reasons like lack of time. Thus, we chunck our test dataset into **3** and save results.
- After preparing the chosen in the format of the model, We got to the **prediction** part where we get our predicted labels at **the token and the line levels**, with the second level being our main interest.

6 Accuracy

- As we got the test predictions in **3 dataframes**, we concatenate them before to get accuracy of the whole DocLayNet base test dataset.
- The following Bar chart is a visualization of the Accuracy by label of the labeled lines of the test dataset



7 Deployment of the model

- Since we worked with HuggingFace from step zero, we opted for a deployment using the **Space** feature of this website, using Gradio as an interface application.
- To get an external link to our Gradio app that lasts for 72h, we resorted to Colab for the deployment code.

7.1 Conclusion

- In this report, we presented how powerful can LiLT(3) get when combined with XLM-RoBERTa(5), a multilingual RoBERTa(6) model (trained on 100 languages) in structured documents layout and understanding. It did a good job when predicting the labels at a line level. Also, during our research, we found the approach of conducting a similar finetuning but on a paragraph level which we will be eager to explore, as well as the intriguing Layout2graph(2) approach that is a language-independent GNN framework for document layout analysis tasks. language scenarios, in addition to the VGT(1) which is a two-stream multi-modal Vision Grid Transformer for document layout analysis and lastly, DiT (Document Image Transformer)(4) that is a self-supervised pre-trained Document Image Transformer model using large-scale unlabeled text images for Document AI tasks..

References

[1] Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao (2023) *Vision Grid Transformer for Document Layout Analysis*

[2] Shu Wei Nuo Xu Deng Huang Xiang Gao (2023) *PARAGRAPH2GRAPH: A GNN-BASED FRAMEWORK FOR LAYOUT PARAGRAPH ANALYSIS*

[3] Jiapeng Wang, Lianwen Jin, Kai Ding (2022) *LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding*

[4] Junlong Li Lei Cui Yiheng Xu Cha Zhang Furu Wei (2022) *DiT: Self-supervised Pre-training for Document Image Transformer*

[5] Naman Goyal Vishrav Chaudhary Guillaume Wenzek Francisco Guzman Edouard Grave Myle Ott LukeZettlemoyer Veselin Stoyanov (2020) *Unsupervised Cross-lingual Representation Learning at Scale*

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov (2019) *RoBERTa: ARobustly Optimized BERT Pretraining Approach*

[7] Minjoon Seo Aniruddha Kembhavi Ali Farhadi Hananneh Hajishirzi (2018) *BI-DIRECTIONAL ATTENTION FLOW FOR MACHINE COMPREHENSION*