SCHOOL OF COMPUTER SCIENCE

Few-shot Document Layout Analysis on Multilingual CFT Documents

Zixuan Xia, Quanxi Li, Jelim Lee
December 19, 2024

1 INTRODUCTION

Document layout analysis involves automatically understanding and categorizing the layout and structure of a document. The ultimate goal is to determine the physical structure of a document, such as headers, paragraphs, tables, and other components. This task is crucial for tasks like digitizing paper documents or extracting meaningful information efficiently.

Various algorithms have been proposed to solve this algorithm. For example, Zhu et al. [1] proposed a novel OCR approach based on document layout analysis. In addition, with the development of deep learning, which includes many labeled samples for training and evaluating neural networks [2], long-term efforts have been made to revolutionize document layout analysis by leveraging neural networks to effectively capture hierarchical and contextual relationships within document layouts. Furthermore, we can roughly divide these methods into two categories, image-centric methods and text-centric methods.

Image-centric approaches, relying solely on visual features, struggle to generalize across diverse document types. Similarly, text-centric methods overlook crucial layout information, making them inadequate for tasks requiring multimodal understanding. These limitations highlight the need for methods that effectively combine textual, visual, and spatial elements.

In recent years, multimodal pre-training for visually-rich Document Understanding (VrDU) has set new state-of-the-art performance on various public benchmarks [3]. These include tasks such as form understanding [4], receipt understanding [5], complex layout understand-

ing, document image classification [6], and document visual question answering [7]. The key advantage lies in the ability to jointly learn text, layout, and image information end-to-end within a unified framework.

Despite these advances, most existing methods heavily rely on large-scale annotated data, which is often unavailable in real-world scenarios. This limitation underscores the importance of few-shot learning techniques, which aim to achieve robust performance with minimal labeled data. Our project distinguishes itself by focusing specifically on few-shot document layout analysis, making it highly applicable to scenarios where annotated data is scarce. Furthermore, our work emphasizes testing the cross-language transfer capabilities of models. By targeting multilingual Calls for Tenders (CFT) documents in German, French, and Italian, our approach provides a rigorous benchmark for evaluating the adaptability and robustness of state-of-the-art models across languages.

2 Related Work

2.1 CNN-based Document Layout Analysis

Convolutional Neural Networks (CNNs) have been widely adopted in document layout analysis, particularly in object detection tasks. This evolution began with the seminal work of Fast R-CNN, which was proposed by Ren et al. [8]. Moreover, Mask R-CNN [9] extended the capabilities by integrating instance segmentation, allowing for precise pixel-level localization of document elements. These methods were transferred to the document layout analysis task [10]. Furthermore, the YOLO (You Only Look Once) family of algorithms has emerged as a prominent framework for real-time object detection, combining speed and accuracy. YOLOv10 [11], the latest iteration, builds on these principles with enhancements in feature extraction and multiscale object detection. This model's properties make it a compelling choice for integration into our workflow.

2.2 Transformer-based Document Layout Analysis

Transformer-based models, known for their positional embedding and attention mechanisms [12], have revolutionized document layout analysis by leveraging self-attention mechanisms to capture long-range dependencies and relationships within documents. In addition, the LayoutLM family of models, including LayoutLM and LayoutLMv2 [13] [14], represented a significant advancement in this domain by integrating textual, visual, and spatial information into a unified framework. Despite these advancements, challenges remain in handling multilingual documents, data scarcity, and adapting models to real-world, low-resource scenarios. LayoutXLM [15] extended these advancements to the multilingual domain, allowing the model to handle documents in multiple languages simultaneously, which demonstrated exceptional performance in cross-lingual transfer tasks, making it highly suitable for analyzing

multilingual Calls for Tenders (CFT) documents.

2.3 Datasets for Document Layout Analysis

Datasets play a critical role in advancing document layout analysis by providing benchmarks for evaluating and training models. Well-known datasets, such as FUNSD [4] and DocVQA [7], focuses on noisy, real-world scanned documents, providing annotations for semantic entities and relationships. These datasets exemplify the challenges inherent in document layout analysis, from handling clean, structured layouts to noisy, multilingual, and complex formats. Our project utilizes PubLayNet [16], a large multilingual public dataset, for pretraining and focuses on multilingual CFT documents during inference, addressing the gap in cross-lingual, few-shot scenarios.

3 Methodology

Our approach consists of two integral parts, which are (pre)training phase and inference phase. To identify the most crucial features for document layout analysis, we emply both CNN-based and transformer-based pretrained models in the (pre)training phase. While in the inference phase, we need to preprocess the input documents to standardize their formats and prepare them for further analysis. Specifically, we need to uniformly process CFT documents in different formats into PDF format, and then split each page of the PDF into pictures. The high-level procedure of our methods are demonstrated in Figure 1.

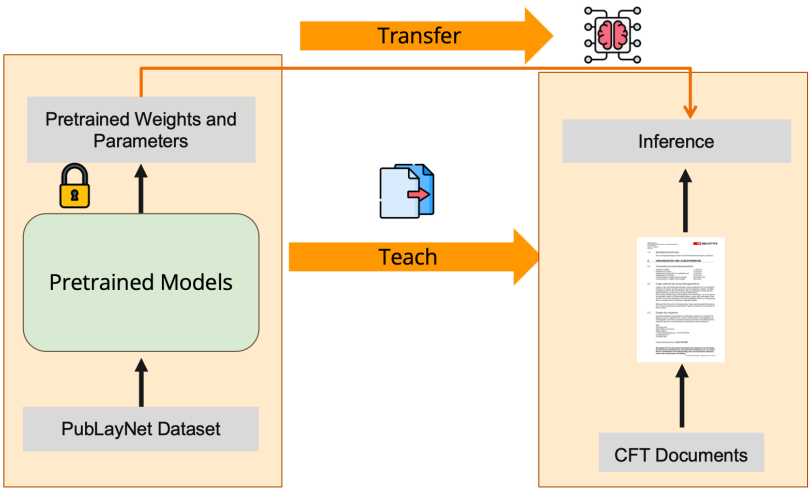

Figure 1: The high-level procedure of our methods

In the (pre)training phase, we first used the YOLOv10 [11] model as the representative of the

CNN-based model, and choosed the LayoutXLM [15] as the representative of the transformer-based model. However, I decided to add the LiLT [17] model, which is a simpler and more effective transformer based model. Unlike LayoutXLM, which emphasizes multilingual language understanding, LiLT focuses on layout structure and the visual aspects of documents. This addition ensures a more robust and versatile approach to document layout analysis.

4 Experiment Setup and Results

Our proposed methodology for few-shot learning demonstrates effectiveness in scenrios with limited labeled data. For each model, we analyzed 10 representative PDF files per language, ensuring consistent use of the same PDF images across different models for a given language to maintain comparability.

Given that our evaluation relies on qualitative methods without predefined ground-truth results, the reported accuracy serves as a reference indicator.The comparative experiments are structed along three key dimensions:

1. **Same model, different languages**
2. **Same language, different models**
3. **Out-of-domain document (For transferability)**

To facilitate subsequent result analysis, we first present all comparison results in Table 1, as shown below.

Table 1: Comparison of Accuracy and Average Inference Time

Model	Acc in DE(%)	Acc in FR(%)	Accuracy in IT(%)	Avg Inference Time (per epoch)
YOLO-v10	70.6	–	–	2.0s
LayoutXLM384	68.3	81.6	84.8%	8.0s
LayoutXLM512	83.1	78.4	76.8	9.0s
LiLT384	81.5	–	–	1.5s
LiLT512	73.8	81.8%	66.0%	4.5s

4.1 Different Languages

The evaluation of the same model across different languages revealed significant variations in performance for both LayoutXLM and LiLT. For LayoutXLM, the smaller model (LayoutXLM384) performs better on French and Italian tasks, while the larger model (LayoutXLM512) excels on German tasks, indicating that different languages have varying adaptability to model size. For LiLT, LiLT384 outperforms LiLT512 on the German task, while LiLT512 achieves higher performance on French but performs poorly on Italian. These results indicate that

the complexity and feature distribution of different languages have distinctimpacts on model performance, further highlighting the need for tailored model configuration and structural variations.

4.2 Different Models

In the German task, LayoutXLM512 achieves the best accuracy but at a higher computational cost. LiLT384, on the other hand, strikes a good balance between accuracy and inference speed, making it a more efficient choice. However, for other tasks, such as Italian, the best performance comes from the LayoutXLM384 model. The significant variations in accuracy demonstrated by other models across German, French, and Italian tasks, suggest that LayoutXLM512 provides strong performance in terms of both accuracy and cross-lingual transferability.

Furthermore, from the perspective of table recognition, models that place greater emphasis on image features tend to perform better in identifying table labels. This may be because tables possess more distinctive visual features, or because image recognition treats text as blocks rather than recognizing it word by word. The specific results are shown in Figures 2(a)–(c).

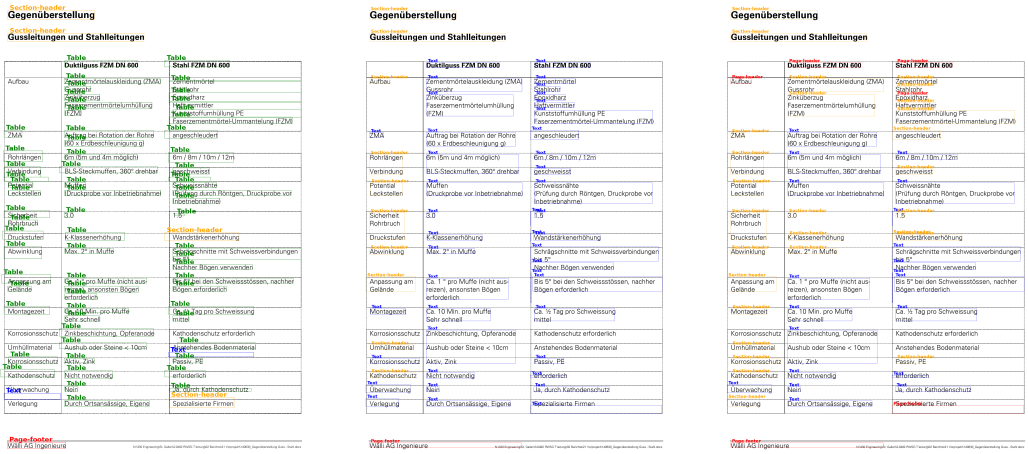

Figure 2: Comparison of table recognition results for different models.

4.3 Noteworth Observations

An intriguing pattern emerges across all models: they predominatly classify character based content as text. This aligns with the conservative and widely applicable assumption that the majority of information in documents can loosely be categorized as text. Moreover, the predominant reading orientation for most contemporary scripts - left-to-right and top-to-bottom-further reinforces this behaviour.This raises an important question: Can these models

effectively recognize documents containing scripts with unconventional reading orders, such as Mongalian, (ancient) Chinese, or Ancient Hebrew?

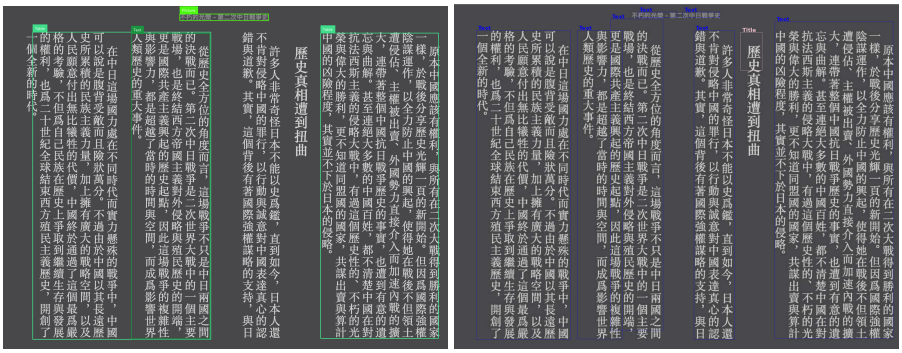

Figure 3: Comparison of ancient Chinese document recognition results for different models.

Figure 3 illustrates a comparative analysis of recognition results for ancient Chinese documents using different models. The YOLOv10 model, for instance, demonstrates a near-complete failure in processing these documents. In contrast, LayoutXLM512 shows marginal success, likely due to its exposure to a limited amount of Chinese data during pretraining. It successfully identifies the document title but tends to classify uncertain regions indiscriminately as text. Although this result may not be ideal, it is worth considering that such document layouts are now rarely used in modern scenarios. Although this result may not be ideal, it is worth considering that such document layouts are now rarely used in modern scenarios. Therefore, we should not expect these models to achieve significant breakthroughs in recognizing these types of documents.

5 Limitations and Future Work

In this study, we relied solely on pretrained models for inference without utilizing private data for training, which limited the model's ability to adapt to domain-specific features. This limitation reduced the models' ability to adapt to the nuanced features of specific domains. For further implementation details, please refer to the public Github Repository [18]. Additionally, the model exhibited instability in few-shot scenarios, as only a small portion of the data was used for inference instead of the entire dataset. This instability highlights the need for improved methods to handle limited data effectively. In future work, we suggest incorporating domain-specific labeled data to further improve accuray and robustness across diverse tasks.

REFERENCES

[1] Weiheng Zhu, Yuanfeng Liu, and Liang Hao. *A Novel OCR Approach Based on Document Layout Analysis and Text Block Classification*. In 2016 12th International Conference on Computational Intelligence and Security (CIS), pages 91–94. IEEE, 2016. doi:10.1109/CIS.2016.0029.

[2] Zejiang Shen, Kaixuan Zhang, and Melissa Dell. *A Large Dataset of Historical Japanese Documents with Complex Layouts*. arXiv preprint arXiv:2004.08686, 2020. Available at: `https://arxiv.org/abs/2004.08686`.

[3] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. *Publaynet: Largest Dataset Ever for Document Layout Analysis*. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1015–1022. IEEE.

[4] Guillaume Jaume, Pablo R. Mendes, and Mathias Niepert. *FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents*. arXiv preprint arXiv:1905.13538, 2019.

[5] Kang Park, Hyunwoo Nam, and Dongkyoo Shin. *Receipt Understanding via Graph Convolutional Networks*. In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[6] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. *Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval*. In Proceedings of the IEEE International Conference on Document Analysis and Recognition (ICDAR), 2015.

[7] Minesh Mathew, Akshay Nagaraja, and Vinay P. Namboodiri. *DocVQA: A Dataset for VQA on Document Images*. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. arXiv preprint arXiv:1506.01497, 2016. Available at: `https://arxiv.org/abs/1506.01497`.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. *Mask R-CNN*. arXiv preprint arXiv:1703.06870, 2018. Available at: `https://arxiv.org/abs/1703.06870`.

[10] Carlos Soto and Shinjae Yoo. *Visual Detection with Context for Document Layout Analysis*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3464–3470, Hong Kong, China, November 2019. Association for Computational Linguistics. Available at: `https://aclanthology.org/D19-1348`. doi:10.18653/v1/D19-1348.

[11] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. *YOLOv10: Real-Time End-to-End Object Detection*. arXiv preprint arXiv:2405.14458, 2024.

Available at: https://arxiv.org/abs/2405.14458.

[12] Tahira Shehzadi, Didier Stricker, and Muhammad Zeshan Afzal. *A Hybrid Approach for Document Layout Analysis in Document Images*. arXiv preprint arXiv:2404.17888, 2024. Available at: https://arxiv.org/abs/2404.17888.

[13] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. *LayoutLM: Pre-training of Text and Layout for Document Image Understanding*. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20), pages 1192–1200, August 2020. ACM. Available at: http://dx.doi.org/10.1145/3394486.3403172. doi:10.1145/3394486.3403172.

[14] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. *LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding*. arXiv preprint arXiv:2012.14740, 2022. Available at: https://arxiv.org/abs/2012.14740.

[15] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. *LayoutXLM: Multimodal Pre-training for Multilingual Visually-Rich Document Understanding*. arXiv preprint arXiv:2104.08836, 2021. Available at: https://arxiv.org/abs/2104.08836.

[16] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. *PubLayNet: Largest Dataset Ever for Document Layout Analysis*. arXiv preprint arXiv:1908.07836, 2019. Available at: https://arxiv.org/abs/1908.07836.

[17] Jiapeng Wang, Lianwen Jin, and Kai Ding. *LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding*. arXiv preprint arXiv:2202.13669, 2022. Available at: https://arxiv.org/abs/2202.13669.

[18] GitHub Repository. *Language Models Repository*. Available at: https://github.com/piegu/language-models/tree/master.