

Multi-Class Document Classification using LayoutLMv1 and V2

Kounen Fathima*, Ali Athar*, Hee-Cheol Kim*

*Department of Digital Anti-Aging Healthcare, Inje University, Gimhae, South Korea

Kounenfathima00@gmail.com, ali.athar1401@gmail.com, heeki@inje.ac.kr

Abstract— In the age of information explosion, efficient and accurate information retrieval has become a pivotal task across numerous domains, from finance to healthcare and beyond. The LayoutLM model has enhanced the capabilities of existing NLP techniques by enabling them to extract information from complexly structured documents automatically. In this study, we employ a subset of the RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset, consisting of 400,000 pictures of data organized into 16 groups. Out of the 16 classes, this subset includes all the classes but there was a limit set to 200 images per class which makes the total amount of images as 3200. Accordingly, 1920, 640, and 640 images make up the training, validation, and testing sets. This dataset is a rich collection of documents with diverse structures and content to demonstrate the effectiveness of our proposed method. LayoutLM, with its unique capability to analyze and understand document structures, has been a pivotal component of our methodology. We utilized LayoutLMv2 and version 1 for the purpose of classifying the documents into their respective categories and comparing the results accordingly. Accordingly, the two versions' accuracies are 80.94 and 68.75 percents.

Keywords— Natural Language Processing (NLP), LayoutLM Model, Document AI, Document Classification, Transformers

I. INTRODUCTION

Digitizing data and extracting information are crucial in modern data centric operations. Business transactions generate much document data, which is unstructured, making extracting useful information challenging. Digitization provides access to cutting-edge data analysis, which can be quickly processed, analyzed, and displayed using advanced tools and algorithms. Information extraction, a crucial step in this process, turns unstructured data into organized, helpful knowledge. Automation is yet another significant advantage. Industry has a practical need for understanding how to digitize data and extract useful information from document data, and academic research on this subject has exploded in recent years.

The two main kinds of documents are scanned images of paper documents and digital documents created by computers, such as PDF, Excel, digital photographs, etc. These documents cover purchasing documents, industry reports, business emails, sales contracts, employment agreements, commercial invoices, personal resumes, and more. Automatically extracting adequate information from these unstructured documents has potential retail value for the development of enterprises [1].

This paper focuses on fine-tuning the LayoutLM model proposed by Huang, Y. et al [2]. This model specializes in Natural Language Processing (NLP) and is designed to perform text recognition and understanding in documents with complex layouts, such as forms, invoices, and receipts [2]. In recent years, deep learning techniques have revolutionized the field, enabling the development of highly accurate and adaptable document classification models. However, documents often exhibit intricate layouts, making it challenging to extract meaningful content for classification accurately. This is especially true in domains where document structure and formatting are pivotal, such as legal documents, financial reports, and medical records. This pre-trained model works on text-centric and image-centric document AI tasks, overcoming the problem. There are various noteworthy uses for this model beyond just categorizing documents. For example, it can also help in understanding documents, comprehending forms, answering visual questions about documents, and understanding receipts. There are other approaches for document classification such as categorizing reviews, tweets or articles with sentiment analysis [3].

This paper will review some related works in section II. Following that, we will delve into the workings of our model in section III. In section IV, we will discuss the results of the model. Finally, in section VI, we will conclude the paper by discussing future work.

II. RELATED WORK

In recent years, document classification has been a well-researched task among others in natural language processing (NLP). Various approaches have been used, including deep learning techniques using CNNs (Convolutional Neural Networks), RNNs (Recurrent Neural Networks), RNTNs (Recursive Neural Tensor Networks), and RCNNs (Recurrent Convolutional Neural Networks) [4].

As more attention is being given to the analysis of visually engaging documents, various approaches have emerged, such as transformer-based, deep learning-based, and hybrid methods. One of the newer hybrid methods combines LayoutLM with BiLSTM and CRF to create a multi-modal fusion technique [1].

Some papers also proposed a hybrid model which works with a new approach which was based on more than just one model like this model named Layout Transformer (LiLT) for

structured document understanding proposed in [5], To obtain the text bounding boxes and contents from an input document image, they made use of commercial OCR engines.

The relevant Transformer-based architecture is then fed with embedded text and layout data to achieve increased functionality. They also made an introduction of the bi-directional attention complementation mechanism (BiACM) which enables the cross-modal interplay of text and layout cues. The pre-training of this model was done on the IIT-CDIP dataset available which consists of 6 million documents and more than 11 million scanned documents [5].

In another research, the paper proposed made a comprehensive analysis between two models namely LayoutLM and Donut. The results showed that the implementation of the LayoutLM performed better with an accuracy of 0.88 than the donut, which obtained 0.74, the comparison was based on the performances of the LayoutLM and Donut in terms of analysis of image classification on different datasets [4].

In paper [6], They proposed a Visual Grid Transformer in which they pre-train the model for 2 Dimensional level for tokens and also semantic understanding at the level of segments. This resulted in great results for the document layout analysis with the best results and the model was named D⁴LA.

There is also research that demonstrates the incorporation of both textual and visual input into a single classification job and pre-training this model using the freely accessible dataset RVL-CDIP. They achieved a high accuracy of 93.03% [7].

More research is available in the betterment of Visual Document Analysis like some more papers proposed different hybrid models like the model named SeRum [8], DocFormer [9] and Donut which is a OCR free document understanding transformer based model [10].

III. METHODOLOGY

In this paper, we intend to understand how document classification works with LayoutLM and analyze its results by comparing the old and newer versions. The LayoutLM proposed by [2] is a simple pretrained language model which works effectively on the layout analysis and text for better document layout understanding. The main applications of this language model are form understanding, information extraction from documents which have complex layout, and receipt understanding. Moreover, the recent version of LayoutLMv3 works by a word-patch alignment aim for predicting if the associated image patch of a text word is hidden while learning cross-modal alignment [11].

A. Data Description

The RVL-CDIP dataset [12], which is freely accessible, comprises document images used to test models for document categorization, information retrieval, and question-answering systems [7]. This dataset was involved in training many models including the LayoutLM model. This paper uses only a small subset of this dataset due to technical limitations.

B. Data Preparation

In this study, text is extracted from photographs and converted into machine-readable English using Tesseract OCR, an open-source optical character recognition (OCR) program. It supports a multitude of languages, making it a versatile tool for text recognition tasks. LayoutLMv2 uses text image alignments and text image matching tasks. It also used the attention mechanism in its transformer architecture which allows the model to understand the relationships between the blocks of text [13].

Hugging Face Transformers Library: Various parameters are used from the transformer library to preprocess the data before the extraction of features. The padding parameters can be used to pad the inputs to a maximum length and ensure that they are all the same length. The truncation parameter can be set to True, which means that if the input data is longer than the specified maximum length, it will be truncated. Label Indexing is done with the help of two dictionaries idx2label and label2idx. Which results in mapping between class labels and their corresponding numerical indices can be done. This indexing is commonly used when preparing data for machine learning and deep learning models, as models work with numerical labels.

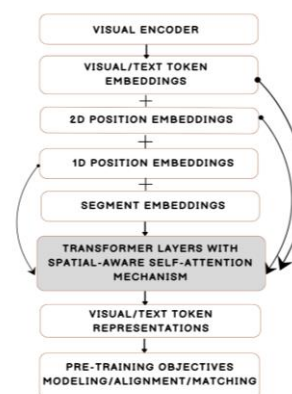


Figure 1. Flow Diagram representing the workflow of LayoutLMv2 [13]

C. Model Architecture

The model basically uses the same architecture in both versions, is that version 2 additionally uses the tesseract OCR open engine and the incorporation of visual embeddings during pre-training in LayoutLMv2. This change suggests an enhancement in leveraging visual information from the document images earlier in the model's training process.

Since the LayoutLM uses transformer architecture and accepts input in the form of image, text, and layout embeddings. It also employs the self-attention technique for improved modelling.

- 1) **Text Embeddings:** These embeddings use models like BERT or RoBERTa to capture semantics and context while encoding textual material.

- 2) **Visual Embeddings:** To comprehend the visual structure of the document, visual aspects like layout and design are retrieved using convolutional neural networks (CNNs).
- 3) **Layout Embeddings:** To aid in exact location comprehension, these embeddings concentrate on spatial information, such as bounding boxes and coordinates.

After merging text, visual, and layout information through embeddings, the model generates feature representation. Depending on the applicable assignment, this full feature representation is subsequently used to either token-level classification or element-level classification. A loss function directs the training process when predictions from the model are contrasted with labels from the ground truth. Once trained, the model is assessed and adjusted as needed to get the best performance. The model is now prepared for inference, which involves analysing newer documents and speculating on their structure and content or categorizing them. This process makes sure that LayoutLMv2 is exceptional in analysing document layouts and associated duties.

Document images use the now trained model for predicting outcomes. The model makes a classification prediction for the document, and the outcomes are converted into class labels and corresponding probabilities. This approach enables the model to categorize the document into the appropriate classifications based on its content.

IV. RESULTS

This experiment evaluates the performances of LayoutLM version 1 and 2. The first version uses the visual embeddings at the stage of fine tuning whereas version 2 adds the embeddings at the time of pre training. The metrics used to calculate the values in table 1 are listed here:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{F1Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

LayoutLM	Precision	Recall	F1-score	Accuracy
Version 1	0.73	0.69	0.69	68.75
Version 2	0.79	0.71	0.75	80.94

Table 1 Comparison of LayoutLM V1 and V2

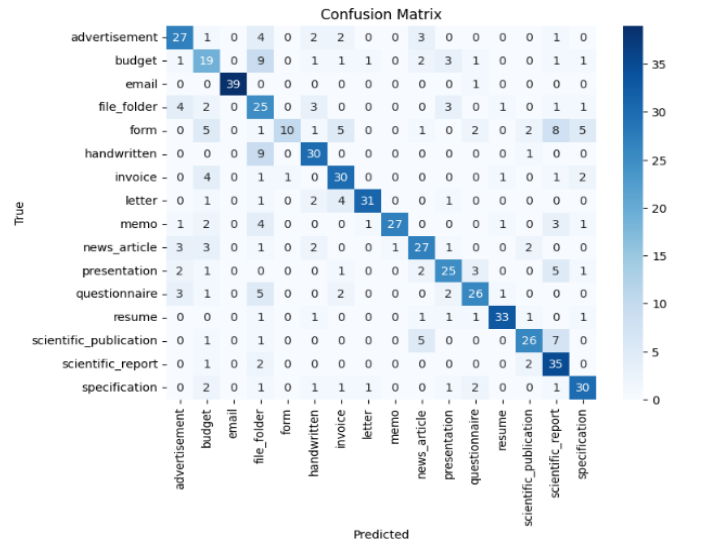


Figure 2. Confusion Matrix for LayoutLMv1

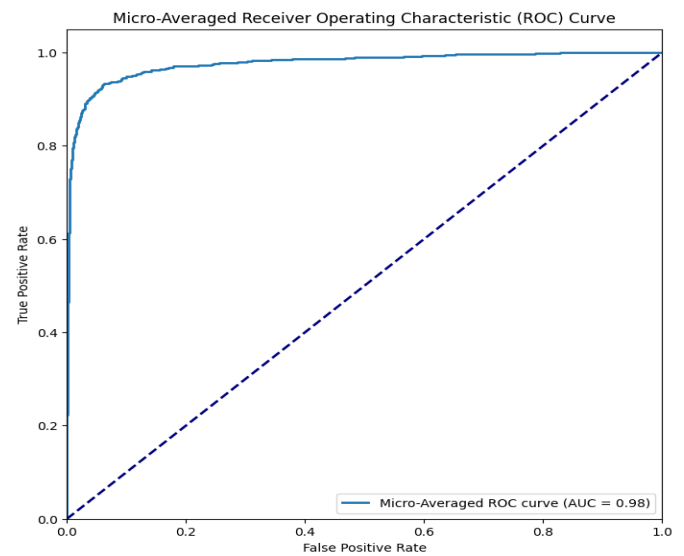


Figure 3. Micro Averaged ROC curve for LayoutLMv2

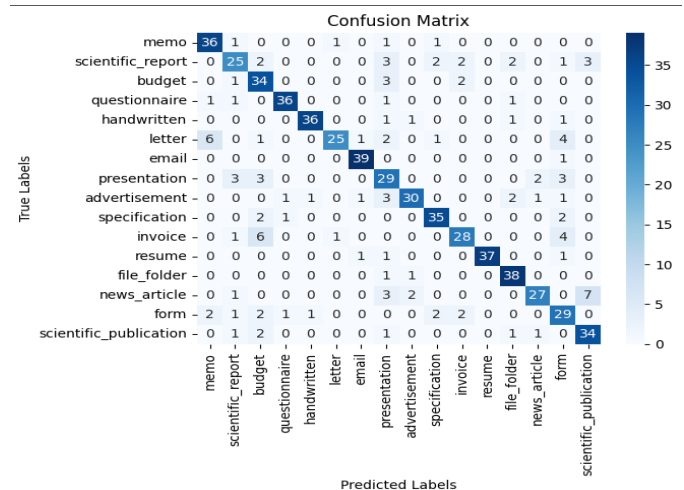


Figure 4. Confusion Matrix for LayoutLMv2

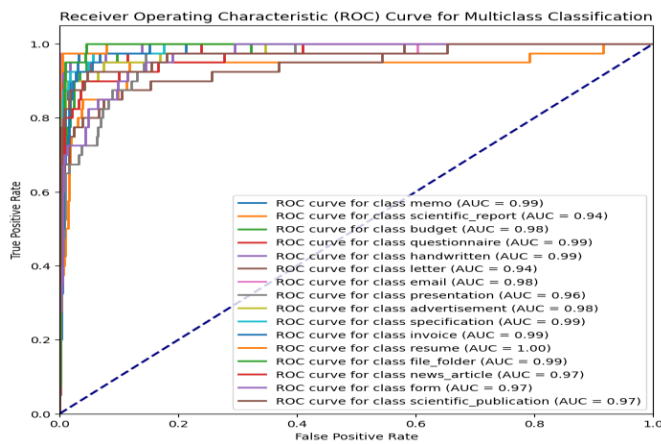


Figure 5 ROC curve for Multiclass classification using LayoutLMv2

In figures 3 and 5, the micro-averaged ROC curve and ROC curve for multiclass classification for all the classes present in the dataset using the LayoutLMv2 can be seen respectively.

V. CONCLUSION AND FUTURE WORK

We employed two versions, which are 1 and 2 of the LayoutLM model and deployed various parameters from the transformer library to prepare the data extracted from the document images. The version 1 of this model achieved an overall accuracy of 68.75% whereas, the Version 2 of this model which uses Tesseract OCR, an open engine to convert the scanned documents into encodings which the machine can read achieved the overall accuracy of 80.94%. We conclude that the LayoutLMv2 model has better performance for many tasks relating to document AI through the VRDs (Visually Rich Documents) compared to the initial version of the same model as it uses not only visual embeddings but also the spatial attributes which results in much better performance in various applications of document AI. Moreover, with the huge trend in digitizing the documents this technology surely has a large room to grow in through creating hybrid approaches to enhance accuracy from it. Future research may use this model's third version and examine how it has improved by utilizing different strategies.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science ICT), Korea, under the National Program for Excellence in SW, supervise by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2022 (2022-0-01091, 1711175863)

REFERENCES

- [1] Boyang Wang and Lin Zhang, "Research on Document Oriented Entity Recognition," *Int. Core J. Eng.*, vol. 8, no. 5, May 2022, doi: 10.6919/ICJE.202205_8(5).0082.
- [2] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, in KDD '20. New York, NY,

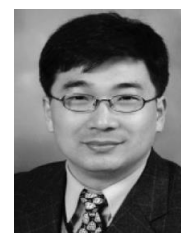
- USA: Association for Computing Machinery, Aug. 2020, pp. 1192–1200. doi: 10.1145/3394486.3403172.
- [3] A. Athar, S. Ali, S. Bhattacharjee, S. A. Shigri, H.-C. Kim, and M. Sheeraz, "Sentimental Analysis of Movie Reviews using Soft Voting Ensemble-based Machine Learning," in *Proceedings of the 2022 IEEE/ACM International Conference on Social Networks and Applied Sciences (SNAMS)*, 2022, doi: 10.1109/SNAMS53716.2021.9732159.
- [4] M. Bajrami, E. Zdravevski, P. Lameski, and B. Stojkoska, "A Comprehensive Analysis of LayoutLM and Donut for Document Classification," *Jul. 2023*, Accessed: Oct. 11, 2023. [Online]. Available: <https://repository.ukim.mk:443/handle/20.500.12188/27397>
- [5] J. Wang, L. Jin, and K. Ding, "LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding," *arXiv*, Feb. 28, 2022, doi: 10.48550/arXiv.2202.13669
- [6] C. Da, C. Luo, Q. Zheng, and C. Yao, "Vision Grid Transformer for Document Layout Analysis," *arXiv*, Aug. 28, 2023. doi: 10.48550/arXiv.2308.14978.
- [7] T. Dauphinee, N. Patel, and M. Rashidi, "Modular Multimodal Architecture for Document Classification," *arXiv*, Dec. 09, 2019. doi: 10.48550/arXiv.1912.04376.
- [8] H. Cao et al., "Attention Where It Matters: Rethinking Visual Document Understanding with Selective Region Concentration," *arXiv*, Sep. 03, 2023, doi: 10.48550/arXiv.2309.01131.
- [9] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "DocFormer: End-to-End Transformer for Document Understanding," *arXiv*, Sep. 20, 2021, doi: 10.48550/arXiv.2106.11539.
- [10] G. Kim et al., "OCR-free Document Understanding Transformer," *arXiv*, Oct. 06, 2022, doi: 10.48550/arXiv.2111.15664.
- [11] "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking | Proceedings of the 30th ACM International Conference on Multimedia." Accessed: Oct. 11, 2023. [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/3503161.3548112>
- [12] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, 2015, pp. 991–995, doi: 10.1109/ICDAR.2015.7333910.
- [13] Y. Xu et al., "LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding," *arXiv*, Jan. 09, 2022, doi: 10.48550/arXiv.2012.14740.



Kounen Fathima received her BE in IT Engineering from the Osmania University, India. Currently, she is pursuing her master's degree in the Institute of Digital Anti-Aging Healthcare from Inje University. Her research interest aligns with Artificial Intelligence, Machine learning and Natural Language Processing.



Ali Athar received his BSSE degree Software Engineering from Government College University Faisalabad (GCUF), Pakistan. He received his MS degree from NUST, Pakistan. He is pursuing his Ph.D. degree from the Institute of Digital Anti-aging and healthcare at Inje University. His research areas include Text Mining, Machine learning, and Deep learning.



Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science in Sogang University in Korea, and Ph.D. at Numerical Analysis and Computing Science, Stockholm University in Sweden in 2001. He is a Professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare, Inje University in Korea. His research interests include machine learning, deep learning, Computer vision, and medical informatics.