Redefining Information Extraction from Visually Rich Documents as Token Classification

Jonghyun Song[1] , **Eunyi Lyou**[1]
[1]Graduate School of Data Science, Seoul National University
{hyeongoon11, onlyou0416}@snu.ac.kr

Abstract

Extracting and retrieving information from visually rich form documents is challenging as it requires considering both text and visual information. The 2024 VRDIU Challenge Track A is focused on improving the performance of bounding box predictions given a document image and a query. To tackle this, we framed the information extraction task as a classification of text tokens. Our model leveraged the cutting-edge performance of LayoutLMv3 by incorporating a token classifier on top of it. We also showed that keeping pixel sizes and aspect ratios close to the original image enhances performance, emphasizing the importance of preventing distortion of visually rich documents. The code for this project is available at THIS LINK.

1 Introduction

Extracting and retrieving information from visually rich form documents is particularly challenging due to the need to comprehend both semantic and visual cues. Often, answers cannot be determined solely by text. For example, the names "Google" and "Tesla" might both be answers to the queries "company name" and "substantial holder name," requiring additional visual information, such as the positions of bounding boxes, for accurate extraction.

To tackle this challenge, we fine-tuned LayoutLMv3 [Huang *et al.*, 2022], a multimodal pretrained transformer that integrates both text and visual data. By leveraging better representation from both semantic and visual cues, we framed the information extraction task as classifying text tokens. Furthermore, our findings show that maintaining pixel sizes close to the original image significantly enhances performance, underscoring the importance of presenting undistorted visual information to the multimodal transformer for understanding visually rich form documents.

2 Methods

2.1 Task and Datasets

VRDIU Challenge (Track A) [ADNLP, 2024] involves the task of accurately locating the Regions of Interest (RoIs) within a document that contain the information required to answer a given query. These RoIs are accompanied by meta-information such as text, images, and positions.

The provided dataset, Form-NLU [Ding *et al.*, 2023], is sourced from the text records of substantial shareholder notice forms submitted to the Australian Stock Exchange. The dataset includes digital, printed, and handwritten images, adding diversity to the data distribution and making the task more challenging. While the training set only consists of digital images, The leaderboard testing set comprises 73% for public and 27% for private, including the printed documents. Therefore, robustness for both digital and printed versions is important for the competition.

2.2 Data preprocessing

We adopted the data preprocessing procedure from LayoutLMv3 to maximize its document understanding capabilities. Our visual input consists of non-overlapping patches that undergo linear projection, inspired by ViT ([Dosovitskiy *et al.*, 2021]). Following LayoutLMv3, we incorporated 1D and 2D positional embeddings to capture spatial information. Additionally, texts extracted from parsed inputs (bounding boxes and corresponding text) for each image is encoded using word embeddings, which are also enhanced with 1D and 2D positional embeddings, in line with the LayoutLMv3 framework. The official implementation of LayoutLMv3 processes images at a resolution of 224 by 224 pixels. However, prior work suggests that preserving the original pixel ratio of images yields better results [Lee *et al.*, 2023]. Therefore, we retained the original resolution as much as possible, setting the image dimensions to 600 by 800 pixels.

2.3 Model Architecture

As the number of queries at the downstream task is limited and predefined by the form designer [Ding *et al.*, 2023], we frame the problem as token-level classification. Specifically, we mapped each query to one of 12 corresponding classes and added a NULL class to represent tokens that do not belong to any query, resulting in a total of 13 classes. Furthermore, since the dataset does not include instances where multiple tokens belong to the same class, we addressed potential overlaps in the model's predictions by processing the evaluation results to ensure that each class prediction remains non-conflicting. For example, the model often confuse the query 'company names' with 'substantial holder names', as

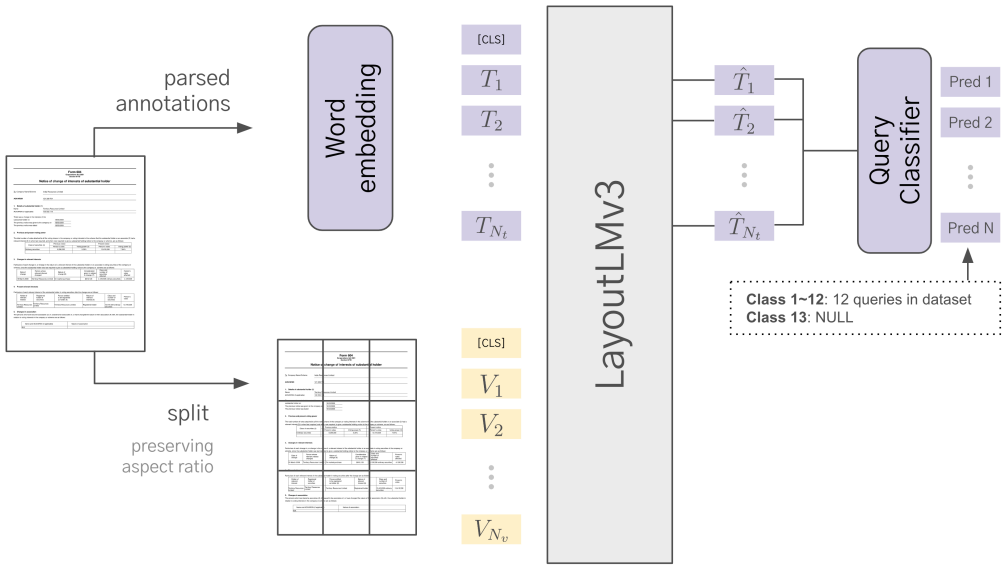

Figure 1: The architecture of LayoutLMv3 fine-tuned for information extraction tasks.

any company names can be mapped to each query. Since the dataset has a fixed format where the company name always precedes the holder name, we used a heuristic to assign "company name" to the earlier token and "substantial holder name" to the later one.

3 Experimental Results

3.1 Training Details

We fine-tune LayoutLMv3-large on Huggingface. The optimization objective is cross-entropy loss with the AdamW optimizer and a learning rate of 2e-6. The model was trained for 10K steps, with the best checkpoint determined by the validation F1 score. Training 10K steps takes approximately 10 hours on a single NVIDIA RTX 4090 GPU.

3.2 Experimental Results

We investigate the performance changes on the leaderboard by varying the training steps and document image resolution, which is represented in Table 1. We explored how the varying number of training steps and the resolution of document images affected performance on the public leaderboard. The results of this investigation are shown in Table 1.

We found that, regardless of the number in training steps, maintaining a resolution close to the original aspect ratio significantly improves performance on the public dataset. Additionally, on private datasets, experiments using the square ratio with which the original LayoutLMv3 was pre-trained performed better than on public datasets. We speculate that the pre-trained weights with the original square ratio are more suitable and robust when dealing with the shift in data distribution from digital training data to printed data in the private dataset. Overall, the minimal performance difference between public and private distributions demonstrates that our method is robust in understanding various types of document data.

Model	Steps	Resolution	public	private
LayoutLMv3	10K	(224, 224)	96.55	97.75
		(600, 800)	97.60	**97.93**
	100K	(224, 224)	96.02	97.75
		(600, 800)	**97.77**	96.72
GPT-3.5-turbo	-	-	31.77	38.28

Table 1: Result of fine-tuned LayoutLMv3 on the public and private leaderboard. Weighted F1 is used as metric.

3.3 Inference with GPT-3.5

Instead of fine-tuning a multi-modal transformer for this task, we analyzed the performance of GPT-3.5-turbo [Brown *et al.*, 2020] when only text information of bounding boxes is given. As shown in Table 1, it was found that GPT-3.5-turbo does not perform well, implying that the form understanding capability of large language models is not well developed yet. Example prompt with one-shot chain-of-thought prompting [Wei *et al.*, 2022] is presented in Appendix.

4 Conclusion

In this work, we presented a method for improving information extraction from visually rich form documents by framing the task as a token classification problem. By fine-tuning LayoutLMv3 with a token classifier, we were able to leverage both text and visual data effectively.

Future work will focus on further optimizing the model and exploring the potential of large language models like GPT for form understanding tasks, as initial results suggest that current models may not yet be fully capable of handling the complexity of these tasks.

References

[ADNLP, 2024] Yihao Ding ADNLP, VRD IU Competition. Vrdiu-track a. https://kaggle.com/competitions/vrd-iu2024-tracka, 2024. VRD IU Competition.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Ding *et al.*, 2023] Yihao Ding, Siqu Long, Jiabin Huang, Kaixuan Ren, Xingxiang Luo, Hyunsuk Chung, and Soyeon Caren Han. Form-nlu: Dataset for the form natural language understanding. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2807–2816, 2023.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Huang *et al.*, 2022] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.

[Lee *et al.*, 2023] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Appendix: Text Prompts for GPT-3.5

Given the query company name, extract the answer object for the query. The query and document are sourced from publicly available financial forms, specifically Form 604 (Notice of Change of Interests of Substantial Holder).
Query is the intended question of the form designer, and the answer is the corresponding user input field. Category 4 and 5 pairs are form query-answer pairs that are horizontally aligned, whereas category 6 and 7 pairs are vertically aligned.
Examples of query and answer text pairs are provided:

- Query: 'Present notice Voting power'
- Answer:
  1. The object that contains the query 'Present notice Voting power' is:
     - global id: 19274, text: Voting Power (5), center x_axis: 431, center y_axis: 497, width: 57, height: 10, category: 6
  2. The possible candidates for the answer objects from the given objects are:
     - global id: 21640, text: 6.04%, center x_axis: 430.0, center y_axis: 508.0, width: 27.0, height: 10.0, category: 7
     - global id: 21641, text: 23,007,197, center x_axis: 345.0, center y_axis: 508.0, width: 46.0, height: 10.0, category: 7
     - global id: 21642, text: 5.03%, center x_axis: 260.0, center y_axis: 508.0, width: 27.0, height: 10.0, category: 7
     - global id: 21643, text: 19,169,682, center x_axis: 175.0, center y_axis: 508.0, width: 46.0, height: 10.0, category: 7
  3. The horizontal distance gaps are:
     - For global id: 21640, distance gap: |431 - 430| = 1
     - For global id: 21641, distance gap: |431 - 345| = 86
     - For global id: 21642, distance gap: |431 - 260| = 171
     - For global id: 21643, distance gap: |431 - 175| = 256
  4. The global id of the final answer is 21640, because it has the smallest distance gap.

Prompt: Given objects from financial form. The answer can only be extracted from this list:

- global id: 18191, text: Form 604 Corporations Act 2001 Section 671B, center x_axis: 264.0, center y_axis: 41.0, width: 85.0, height: 44.0, category: 1
- global id: 18192, text: Notice of change of interests of substantial holder, center x_axis: 169.0, center y_axis: 89.0, width: 274.0, height: 16.0, category: 1
- global id: 18193, text: 1. Details of substantial holder (1), center x_axis: 73.0, center y_axis: 174.0, width: 123.0, height: 11.0, category: 2
- global id: 18194, text: 2. Previous and present voting power, center x_axis: 70.0, center y_axis: 309.0, width: 141.0, height: 13.0, category: 2
- global id: 18195, text: The total number of votes attached to all the voting shares in the company or voting interests in the scheme that the substantial holder or an associate (2) had a relevant interest (3) in when last required, and when now required, to give a substantial holding notice to the company or scheme, are as follows:, center x_axis: 73.0, center y_axis: 329.0, width: 458.0, height: 29.0, category: 3
- global id: 18196, text: Previous notice, center x_axis: 214.0, center y_axis: 366.0, width: 54.0, height: 11.0, category: 3

- global id: 18197, text: Present notice, center x_axis: 381.0, center y_axis: 365.0, width: 53.0, height: 12.0, category: 3
- global id: 18198, text: 3. Changes in relevant interests, center x_axis: 73.0, center y_axis: 455.0, width: 115.0, height: 8.0, category: 2
- global id: 20056, text: To Company Name/Scheme, center x_axis: 72, center y_axis: 126, width: 101, height: 11, category: 4
- global id: 20057, text: ACN/ARSN, center x_axis: 73, center y_axis: 143, width: 42, height: 10, category: 4
- global id: 20058, text: Name, center x_axis: 73, center y_axis: 191, width: 22, height: 8, category: 4
- global id: 20059, text: ACN/ARSN (if applicable), center x_axis: 73, center y_axis: 207, width: 88, height: 10, category: 4
- global id: 20060, text: There was a change in the interests of the substantial holder on, center x_axis: 73, center y_axis: 238, width: 143, height: 17, category: 4
- global id: 20061, text: The previous notice was given to the company on, center x_axis: 73, center y_axis: 260, width: 168, height: 11, category: 4
- global id: 20062, text: The previous notice was dated, center x_axis: 73, center y_axis: 278, width: 105, height: 9, category: 4
- global id: 20063, text: Class of securities (4), center x_axis: 85, center y_axis: 367, width: 74, height: 9, category: 6
- global id: 20064, text: Person’s votes, center x_axis: 214, center y_axis: 382, width: 51, height: 9, category: 6
- global id: 20065, text: Voting power (5), center x_axis: 299, center y_axis: 382, width: 58, height: 10, category: 6
- global id: 20066, text: Person’s votes, center x_axis: 383, center y_axis: 382, width: 51, height: 9, category: 6
- global id: 20067, text: Voting power (5), center x_axis: 456, center y_axis: 382, width: 57, height: 10, category: 6
- global id: 22417, text: 5.15%, center x_axis: 456.0, center y_axis: 397.0, width: 26.0, height: 9.0, category: 7
- global id: 22418, text: 11,473,829, center x_axis: 383.0, center y_axis: 396.0, width: 52.0, height: 10.0, category: 7
- global id: 22419, text: 6.21%, center x_axis: 299.0, center y_axis: 397.0, width: 27.0, height: 9.0, category: 7
- global id: 22420, text: 13,590,540, center x_axis: 214.0, center y_axis: 396.0, width: 54.0, height: 10.0, category: 7
- global id: 22421, text: Ordinary, center x_axis: 83.0, center y_axis: 397.0, width: 44.0, height: 10.0, category: 7
- global id: 22422, text: Avoca Resources Limited, center x_axis: 185.0, center y_axis: 125.0, width: 120.0, height: 11.0, category: 5
- global id: 22423, text: JPMorgan Chase & Co. and its affiliates, center x_axis: 183.0, center y_axis: 189.0, width: 204.0, height: 12.0, category: 5
- global id: 22424, text: N/A, center x_axis: 183.0, center y_axis: 206.0, width: 21.0, height: 11.0, category: 5
- global id: 22425, text: 30/Nov/2007, center x_axis: 253.0, center y_axis: 277.0, width: 58.0, height: 9.0, category: 5
- global id: 22426, text: 30/Nov/2007, center x_axis: 253.0, center y_axis: 260.0, width: 58.0, height: 9.0, category: 5
- global id: 22427, text: 31/Dec/2008, center x_axis: 253.0, center y_axis: 244.0, width: 58.0, height: 10.0, category: 5

Task: extract the object including query 'company name'. Then, find the possible answer objects for the query. If there are multiple candidates for the answer objects, choose the one based on the following rules. Distance gap is **the absolute value** of the difference between two points:

- When the category of query is 4, choose the answer whose category is 5 and which has the closest y_axis (smallest vertical distance gap) to the query.
- When the category of query is 6, choose the answer whose category is 7 and which has the closest x_axis (smallest horizontal distance gap) to the query.

Let's think step by step:

1. List the object that contains the query.
2. List all possible candidates for the answer objects from the given objects. Calculate the distance gap by absolute value between two points.
   - If the answer category is 5, find the vertical (y_axis) distance gap.
   - If the answer category is 7, find the horizontal (x_axis) distance gap.
3. Choose the final answer from candidates with the smallest distance gap in the format of 'the global id of the final answer is [id]'. If the answer is not found, return -1 as id.