LAMPRET: Layout-Aware Multimodal PreTraining for Document Understanding

Te-Lin Wu[*1], **Cheng Li**[2], **Mingyang Zhang**[2], **Tao Chen**[2],
Spurthi Amba Hombaiah[2], **Michael Bendersky**[2]
[1]University of California, Los Angeles, [2]Google Research
telinwu@cs.ucla.edu, {chgli,mingyang,taochen,spurthiah,bemike}@google.com

Abstract

Document layout comprises both structural and visual (*e.g.* font-sizes) information that are vital but often ignored by machine learning models. The few existing models which do use layout information only consider *textual* contents, and overlook the existence of contents in other modalities such as images. Additionally, spatial interactions of presented contents in a layout was never really fully exploited.

To bridge this gap, we parse a document into content blocks (*e.g.* text, table, image) and propose a novel layout-aware multimodal hierarchical framework, LAMPreT, to model the blocks and the whole document. Our LAMPreT encodes each block with a multimodal transformer in the lower-level, and aggregates the block-level representations and connections utilizing a specifically designed transformer at the higher-level. We design hierarchical pretraining objectives where the lower-level model is trained similarly to multimodal grounding models, and the higher-level model is trained with our proposed novel layout-aware objectives. We assess the proposed model on two layout-aware tasks – text block filling and image suggestion, and show the effectiveness of our proposed hierarchical architecture as well as pretraining techniques.

1 Introduction

Layout, the structural and visual presentation of the contents, is a key aspect for composing documents. Specifically, the planning and the arrangements of how the contents are spatially structured, as well as the use of multiple modalities (*e.g.* texts, graphics, tables), is highly influential in the choice of reading strategies from the readers, and hence is vital for document understanding (Hartley, 2013).

Learning a document-level representation with awareness of the layout has started to become an active research area, especially in achieving better semantic document understanding (Katti et al., 2018; Denk and Reisswig, 2019; Xu et al., 2020). However, most of the prior works focus on rather surface forms of the layout (Wang et al., 2020) and the claimed multimodality refers to OCR detected features of the textual components (Zhang et al., 2020; Kerroumi et al., 2020) rather than the actual multimedia contents (*e.g.* images). Moreover, they mostly concern documents in the domain of scanned templates, *e.g.* receipts (Pramanik et al., 2020), and are not as *content-rich* and *flexible* as articles like, *e.g.* Wikipedia pages.

In this paper, we propose **L**ayout-**A**ware **M**ultimodal **PreT**raining (LAMPRET), aiming for a more general-purposed pretraining methodology which exploits both the structure and the content of documents, and considers multimedia contents, such as images, to learn a comprehensive multimodal document representation. Specifically, we utilize an in-house document tokenizer to parse HTML formatted pages into several *content blocks*, where each *block* has the following features: (1) spatial position, (2) semantic types, *e.g.* headers and tables, and (3) attributes, *e.g.* font-sizes.

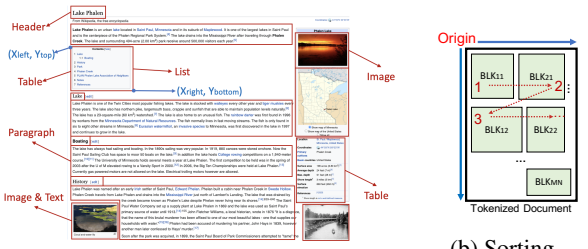
Inspired by the inherent hierarchy in the contents, our LAMPRET framework is hierarchical, consisting of two cascaded transformers (Vaswani et al., 2017). The lower-level transformer takes as inputs the parsed multimodal content blocks serialized by their sorted spatial positions, and the output *block-level* representations from the former are consumed by the higher-level transformer. The lower-level model is trained with the Masked Language Modeling (MLM) objective (Devlin et al., 2019) and an **image-to-text matching prediction** for grounding different input modalities. For training the higher-level model, we propose three novel *block-level* pretraining objectives aiming to exploit the structure of a document: (1) **block-ordering prediction** requires the model to predict whether

[*]Work done during an internship at Google Research.

the input blocks are properly ordered, (2) **masked-block prediction** shares similar spirit with textual MLM but acts at the textual *block-level*, and (3) **image fitting prediction** requires the model to select the most suitable image for a missing image block.

We evaluate our proposed LAMPRET framework on two downstream document completion tasks: (1) **Text block filling** which aims to select the most appropriate textual block for a missing block to complete a document, and (2) **Image content suggestion** where the models are required to correctly retrieve the most suitable image at a layout position for a particular document, from a sizable set of candidates approximating realistic scenarios of composing documents. We show the effectiveness of LAMPRET and the benefits of incorporating multimodality, and we also conduct extensive ablation studies on its components.

2 Document Layout



(a) Parser. (b) Sorting.

Figure 1: **(a) An example page parsed by the document tokenizer.** Each red box indicates a content block. Blue colored coordinates are an exemplar block position tuple. **(b) Sorting and serializing blocks:** Setting the *origin* at the top-left-most corner, a 2D sorting is performed according to the *top-left* block position, and then serialized with a *zigzag* fashion.

Document Tokenizer: In this work, the layout is obtained using an in-house HTML formatted webpage document parsing tool[1]. Figure 1a illustrates how a document is *tokenized* (parsed) into several small *content* ***blocks***, each is a proportion of the document showing a clear spatial boundary to the others. Each block has the following features: **(1) Block Position:** the 2D real valued position of the bounding box which encompasses the block. **(2) Block Type:** the semantic type of the content presented in the block, such as header, paragraph, image, lists, table, etc. **(3) Block Attributes:** the visual presentations of the **texts** featured in a block, such as font-size, and **bold**, *italic*, or underlined. **(4) Multimedia:** in this work, we only consider images as the multimedia contents for our models.

Layout: We define layout as the structural presentation of the tokenized content blocks, *i.e.* their relative positions and orders, and the aforementioned attributed features of the textual contents within a block. We first ***sort*** the tokenized content blocks, with respect to the block positions, and then serialize them in a ***zigzag*** fashion to construct inputs to our models, as illustrated in Figure 1b.

3 LAMPreT

3.1 Model Overview

Hierarchical Architecture: We design a framework consisting of two levels of transformers, where the lower-level concerns contents of a block while the higher level handles how these blocks are spatially structured, as illustrated in Figure 2a. The lower-level model takes as *multimodal* inputs of the parsed contents, where each *content block* is placed at its **serialized sorted** position. Each block (blk_i) is prepended with a CLS_i special token for indicating the boundary of block contents. We also prepend a global-CLS token at the beginning of the inputs for obtaining *document-level* representation. The higher-level model then takes as inputs the block-level representations blkh_i, *i.e.* the outputs of the lower-level model at each CLS_i position.

Input Representations: Apart from the token embedding, the input representation contains the embeddings from: (1) block attributes, (2) a *block-segment-id* per each block, (3) multimodality indicator, and (4) visual inputs where a convolutional neural network (CNN) and transformation multilayer perceptron (MLP) are adopted.

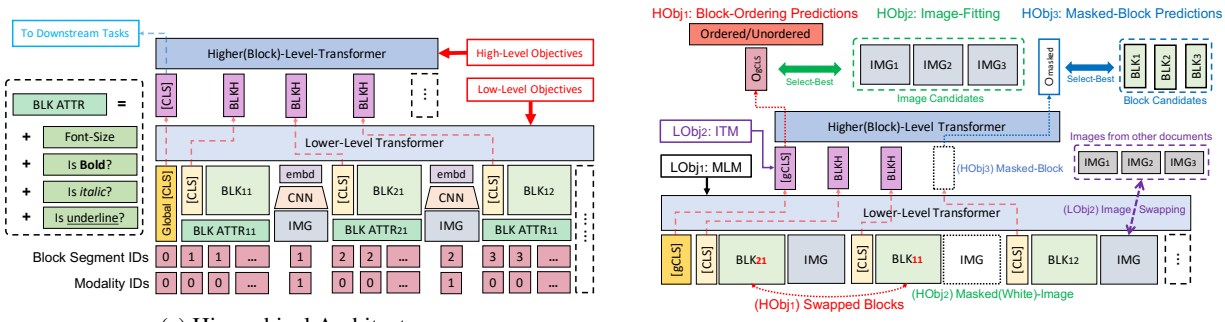
3.2 Training Objectives

Figure 2b shows the overall training objectives of LAMPRET. The lower-level training objectives include: (1) **Masked Language Modeling (MLM)** (Devlin et al., 2019), and (2) **Image-Text Matching (ITM)** that requires the model to predict whether the texts and the images are aligned. Similar to (Lu et al., 2019), images of a document are probabilistically swapped with those in other documents within the same training mini-batch.
The higher-level training objectives include:

- **Block-Ordering Predictions (B-ORD):** Two input blocks are randomly selected and swapped[2]

[1]Note that document parsing is not of our main focus, we assume our data is already parsed by the ready-to-use tool.

[2]We limit the random re-ordering of the blocks to 2 to



(a) Hierarchical Architecture.

(b) Training Objectives.

Figure 2: **(a) Hierarchical formulation:** LAMPRET framework exploits the inherent hierarchical nature of document layouts. The input representation of each block blk_i contains the embeddings of Wordpiece tokens, block-segment-ids, modalities, and attributional features. The output representations of the lower-level model at each CLS_i position are fed to the higher-level model. **(b) Training objectives:** HObj_i and LObj_i denotes the i-th high- and low-level objective respectively. For each high-level objective, we illustrate: an exemplar block swapping for the block-ordering objective, an image masking for the image fitting objective, and a block masked at its *block-level representation* for the block-MLM objective, respectively.

with certain probability in their serialized order when inputting to the the lower-level model. The model is required to make binary *ordered* or *unordered* predictions using the output representation at the global-CLS position.

- **Block-MLM (B-MLM):** One or more textual blocks are **masked out** at their *block-level* representations, blkh_i, for which the model is required to *select* the most suitable block from a given set of candidate blocks within the training mini-batch. We cast it as a classification problem where prediction is made on a concatenation of output representations of the masked blocks and the *block-level* representations of the candidates.
- **Image Fitting (IMG-FIT):** One or more images are masked out, for which the model needs to select the most suitable images from a set of candidates constructed within the training mini-batch. We similarly cast it as a classification problem with the inputs of: **concat**($\text{out}_{global-CLS}$, $\text{out}_{\text{blk},i}$, $\text{embd}_{img,1}$, $\text{embd}_{img,2}$, ...), where $\text{embd}_{img,j}$ is the visual embedding of the j-th image candidate. The output representation at the global-CLS position is incorporated to model the general trends of how the images are positioned in a document.

4 Experiments

4.1 Evaluation Tasks

Text Block Filling: We randomly select a block blk_i to mask within $\{\text{blk}_1, \text{blk}_2, ..., \text{blk}_N\}$, and provide the context $\text{blk}_{1:i-1}$ as inputs to the model, while leaving $\text{blk}_{j:j+K} \cup \text{blk}_i$ as candidates to select, where $j > i$. blk_j is spatially positioned after blk_i by a certain margin[3], and K is set to 5.

Image Suggestion: The model takes as inputs all the content blocks of a document with **an image masked-out**, and is required to predict the correct image from a given set of candidates. We extract $C = 1000$ candidate images from documents unseen during the pretraining for this evaluation.

Finetuning on Downstream Tasks: To allow better fusion of low and high-level information, we have: $R_{doc} = \sigma(\alpha) \cdot \text{blkh}_{global-CLS} + (1 - \sigma(\alpha)) \cdot \text{out}_{global-CLS}$ to represent a document, where α is a learnable scalar and σ is the Sigmoid function. We adopt a contrastive loss (Hadsell et al., 2006) to train R_{doc} with candidate embeddings $\{R_{cand}\}$, where $R_{cand,i} = \text{blkh}_i$ for text block filling and $R_{cand,i} = \text{embd}_{img,i}$ for image suggestion:

$$D_w(R_{doc}, R_{cand,i}) = \|\textbf{MLP}(R_{doc}) - \textbf{MLP}(R_{cand,i})\|_2$$
$$L_{\text{contrastive}} = (1 - Y)\frac{1}{2}(D_w)^2 + (Y)\frac{1}{2}\{max(0, m - D_w)\}^2 \quad (1)$$

4.2 Baselines

Single-Level LayoutLM: The single-level, non-hierarchical variant of LAMPRET framework consists of only the lower-level model, which resembles the base model of the prior work LayoutLM (Xu et al., 2020). For training, both low-level objectives, MLM and ITM, are utilized.

CNN-Grid: Inspired by prior work ChargGrid (Katti et al., 2018), we replace the transformer-higher-level model with a CNN module. Each *block-level* representation blkh_i is inserted to a 2D

leave majority of other blocks untouched for other objectives.

[3]Four rows of normal text in Wikipedia pages.

Method	Modality	Text Block Filling			Image Suggestion (C=1000)	
		F1-Score	Prec. (%)	Rec. (%)	Rec. @ 5 (%)	MRR (%)
CNN-Grid	Multimodal	39.92	40.45	39.40	67.33	62.60
Single-Level LayoutLM	Text-Only	51.49	39.93	**72.45**	—	—
	Multimodal	51.30	41.40	67.43	75.99	76.54
LAMPRET	Text-Only	**52.36***	**42.37***	68.50	—	—
	Multimodal	52.09*	41.85*	68.98	**99.98***	**98.55***

(a) Model Performances.

Ablation	Text Block Filling	Image Suggestion
	F1-Score	MRR (%)
w/o IMG-FIT	50.72	89.68
w/o B-MLM	51.18	85.26
w/o B-ORD	50.19	69.51
w/o Attributes	50.98	95.27
Multimodal LAMPRET	52.09	98.55

(b) LAMPRET Ablations.

Table 1: **(a) Model Performances:** Best performances for each metric is boldfaced. Our LAMPRET framework outperforms the carefully crafted baseline models in both tasks. * *indicates that our* LAMPRET *framework is statistically significantly better than the best-performing baselines, according to the size of our test-set at a level of significance of 0.01.* **(b) Model Ablation Studies:** We examine the contribution of the high-level pretraining objectives on the multimodal version of LAMPRET. Each row denotes the pretraining conducted with the indicated objective excluded, additionally with the last row without the attributed features, which are proven more effective on the text-based task. We train for much more iterations during downstream finetuning for the image suggestion task (until performance convergence) when the model is not pretrained with the image fitting objective.

position of a 3D map according to the sorted coordinates, where the original 1D blkh_i representation becomes the channel dimension. An average pooling at the output of the CNN is adopted to obtain document-level representation, and the same set of objectives in LAMPRET are applied for training.

Unimodality (Text-Only): We experiment with the text-only variant of models for both LAMPRET and the single-level LayoutLM, for comparisons particularly for the text block filling task.

4.3 Experimental Setups

Dataset We scrape a collection of English Wikipedia pages for training and evaluating the models. Our dataset is uniformly scraped from the entire Wikipedia, resulting in a total of 6M pages.

Evaluation Metrics As our tasks are of retrieval fashion, we adopt two ranking-based metrics: (1) Mean Reciprocal Rank, and (2) Recall@K.

4.4 Experimental Results

Table 1a summarizes the model performances.

Text Block Filling. For the text block filling downstream task, our proposed LAMPRET model outperforms the baselines in both the precision and F-1 score metrics. It is worth noting that for both LAMPRET and the single-level LayoutLM, the unimodal text-only version performs slightly better than the multimodal version. We hypothesize that such results can be attributed to the suboptimal multimodal representation fusing, that it can be potentially alleviated with more sophisticated and finer-grained multimodal grounding paradigms. Among the models, CNN-Grid baselines performs the worst, of which the attention mechanism in transformers is hypothesized to capture the block-level interactions better.

Image Suggestion. This task is only evaluated on multimodal models, as unimodal ones do not have access to image based features. Our LAMPRET achieves almost perfect performance for both metrics (99%), while all the baseline models suffer significant performance degradation. The hierarchical formulation and the accompanying high-level objectives are proven effective in LAMPRET as compared to single-level LayoutLM. The CNN-Grid baseline again generally performs the worst on both metrics.

Model Ablation Studies. Table 1b shows an ablation analysis of the multimodal version of LAMPRET on the contributions of the pretraining objectives for different downstream tasks. At each row, we exclude (1) one of the high-level pretraining objectives, or (2) the layout attributed features. In general, the block-ordering objective is empirically proven quite effective for both downstream tasks, judged by the performance degradation when it is excluded. It is worth noting as well that the exclusion of the layout attributes do not cause much deterioration compared to other pretraining objectives, which hypothetically implies that the exploitation of the structural information of layout designed in LAMPRETis relatively more effective.

5 Conclusions

We propose a multimodal layout-aware document representation learning framework, LAMPRET. Once a document is parsed into several spatially structured *content blocks*, we sort and serialize them in a 2D formulation. LAMPRET models the inherent hierarchical formulation of a document layout using two cascaded transformers, where the lower-level is trained with MLM and ITM objectives, while the higher-level is trained with three specifically designed layout-exploiting objectives. We evaluate LAMPRET on two downstream tasks: (1) text block filling, and (2) image suggestion task.

References

Timo I Denk and Christian Reisswig. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. In *Workshop on Document Intelligence at NeurIPS 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

James Hartley. 2013. *Designing instructional text*. Routledge.

Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4459–4469.

Mohamed Kerroumi, Othmane Sayem, and Aymen Shabou. 2020. Visualwordgrid: Information extraction from scanned documents using a multimodal approach. *arXiv preprint arXiv:2010.02358*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23.

Subhojeet Pramanik, Shashank Mujumdar, and Hima Patel. 2020. Towards a multi-modal, multi-task learning based pre-training framework for document representation learning. *arXiv preprint arXiv:2009.14457*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. 2020. Docstruct: A multimodal method to extract hierarchy structure in document for general form understanding. In *Findings of Empirical Methods in Natural Language Processing (EMNLP)*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pretraining of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. 2020. Trie: End-to-end text reading and information extraction for document understanding. In *ACM International Conference on Multimedia (ACMMM)*.