

LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking

Yupan Huang*
Sun Yat-sen University
huangyp28@mail2.sysu.edu.cn

Tengchao Lv
Microsoft Research Asia
tengchaolv@microsoft.com

Lei Cui
Microsoft Research Asia
lecu@microsoft.com

Yutong Lu
Sun Yat-sen University
luyutong@mail.sysu.edu.cn

Furu Wei
Microsoft Research Asia
fuwei@microsoft.com

ABSTRACT

Self-supervised pre-training techniques have achieved remarkable progress in Document AI. Most multimodal pre-trained models use a masked language modeling objective to learn bidirectional representations on the text modality, but they differ in pre-training objectives for the image modality. This discrepancy adds difficulty to multimodal representation learning. In this paper, we propose **LayoutLMv3** to pre-train multimodal Transformers for Document AI with unified text and image masking. Additionally, LayoutLMv3 is pre-trained with a word-patch alignment objective to learn cross-modal alignment by predicting whether the corresponding image patch of a text word is masked. The simple unified architecture and training objectives make LayoutLMv3 a general-purpose pre-trained model for both text-centric and image-centric Document AI tasks. Experimental results show that LayoutLMv3 achieves state-of-the-art performance not only in text-centric tasks, including form understanding, receipt understanding, and document visual question answering, but also in image-centric tasks such as document image classification and document layout analysis. The code and models are publicly available at <https://aka.ms/layoutlmv3>.

CCS CONCEPTS

• **Applied computing** → **Document analysis**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

document ai, layoutlm, multimodal pre-training, vision-and-language

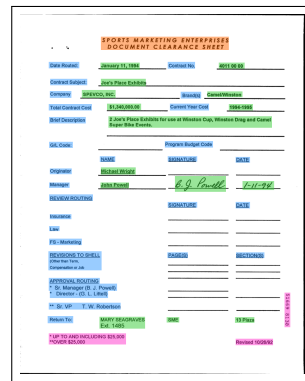
ACM Reference Format:

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3503161.3548112>

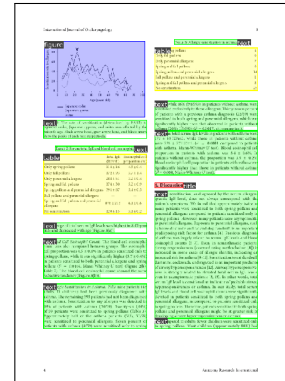
*Contribution during internship at Microsoft Research. Corresponding authors: Lei Cui and Furu Wei.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions.acm.org.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10...\$15.00
<https://doi.org/10.1145/3503161.3548112>



(a) Text-centric form understanding on FUNSD



(b) Image-centric layout analysis on PubLayNet

Figure 1: Examples of Document AI Tasks.

1 INTRODUCTION

In recent years, pre-training techniques have been making waves in the Document AI community by achieving remarkable progress on document understanding tasks [2, 13–15, 17, 25, 28, 31, 32, 40, 41, 50, 52, 54–56]. As shown in Figure 1, a pre-trained Document AI model can parse layout and extract key information for various documents such as scanned forms and academic papers, which is important for industrial applications and academic research [8].

Self-supervised pre-training techniques have made rapid progress in representation learning due to their successful applications of reconstructive pre-training objectives. In NLP research, BERT firstly proposed “masked language modeling” (MLM) to learn bidirectional representations by predicting the original vocabulary id of a randomly masked word token based on its context [9]. Whereas most performant multimodal pre-trained Document AI models use the MLM proposed by BERT for text modality, they differ in pre-training objectives for image modality as depicted in Figure 2. For example, DocFormer learns to reconstruct image pixels through a CNN decoder [2], which tends to learn noisy details rather than high-level structures such as document layouts [43, 45]. SelfDoc proposes to regress masked region features [31], which is noisier and harder to learn than classifying discrete features in a smaller vocabulary [6, 18]. The different granularities of image (e.g., dense image pixels or contiguous region features) and text (i.e., discrete

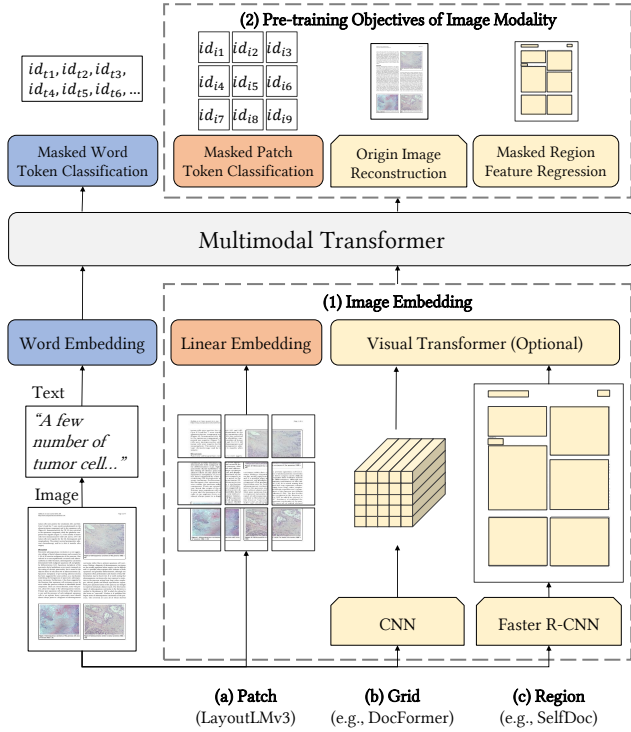


Figure 2: Comparisons with existing works (e.g., DocFormer [2] and SelfDoc [31]) on (1) image embedding: our LayoutLMv3 uses linear patches to reduce the computational bottleneck of CNNs and eliminate the need for region supervision in training object detectors; (2) pre-training objectives on image modality: our LayoutLMv3 learns to reconstruct discrete image tokens of masked patches instead of raw pixels or region features to capture high-level layout structures rather than noisy details.

tokens) objectives further add difficulty to cross-modal alignment learning, which is essential to multimodal representation learning.

To overcome the discrepancy in pre-training objectives of text and image modalities and facilitate multimodal representation learning, we propose **LayoutLMv3** to pre-train multimodal Transformers for Document AI with unified text and image masking objectives MLM and MIM. As shown in Figure 3, LayoutLMv3 learns to reconstruct masked word tokens of the text modality and symmetrically reconstruct masked patch tokens of the image modality. **Inspired by DALL-E [43] and BEiT [3], we obtain the target image tokens from latent codes of a discrete VAE.** For documents, each text word corresponds to an image patch. To learn this cross-modal alignment, we propose a Word-Patch Alignment (WPA) objective to predict whether the corresponding image patch of a text word is masked.

Inspired by ViT [11] and ViLT [22], LayoutLMv3 directly leverages raw image patches from document images without complex pre-processing steps such as page object detection. LayoutLMv3 jointly learns image, text and multimodal representations in a Transformer model with unified MLM, MIM and WPA objectives. This makes LayoutLMv3 the first multimodal pre-trained Document AI

model without CNNs for image embeddings, which significantly saves parameters and gets rid of region annotations. The simple unified architecture and objectives make LayoutLMv3 a general-purpose pre-trained model for both text-centric tasks and image-centric Document AI tasks.

We evaluated pre-trained LayoutLMv3 models across five public benchmarks, including text-centric benchmarks: FUNSD [20] for form understanding, CORD [39] for receipt understanding, DocVQA [38] for document visual question answering, and image-centric benchmarks: RVL-CDIP [16] for document image classification, PubLayNet [59] for document layout analysis. Experiment results demonstrate that LayoutLMv3 achieves state-of-the-art performance on these benchmarks with parameter efficiency. Furthermore, LayoutLMv3 is easy to reproduce for its simple and neat architecture and pre-training objectives.

Our contributions are summarized as follows:

- LayoutLMv3 is the first multimodal model in Document AI that does not rely on a pre-trained CNN or Faster R-CNN backbone to extract visual features, which significantly saves parameters and eliminates region annotations.
- LayoutLMv3 mitigates the discrepancy between text and image multimodal representation learning with unified discrete token reconstructive objectives MLM and MIM. We further propose a Word-Patch Alignment (WPA) objective to facilitate cross-modal alignment learning.
- LayoutLMv3 is a general-purpose model for both text-centric and image-centric Document AI tasks. For the first time, we demonstrate the generality of multimodal Transformers to vision tasks in Document AI.
- Experimental results show that LayoutLMv3 achieves state-of-the-art performance in text-centric tasks and image-centric tasks in Document AI. The code and models are publicly available at <https://aka.ms/layoutlmv3>.

2 LAYOUTLMV3

Figure 3 gives an overview of the LayoutLMv3.

2.1 Model Architecture

LayoutLMv3 applies a unified text-image multimodal Transformer to learn cross-modal representations. The Transformer has a multi-layer architecture and each layer mainly consists of multi-head self-attention and position-wise fully connected feed-forward networks [49]. The input of Transformer is a concatenation of text embedding $Y = y_{1:L}$ and image embedding $X = x_{1:M}$ sequences, where L and M are sequence lengths for text and image respectively. Through the Transformer, the last layer outputs text-and-image contextual representations.

Text Embedding. Text embedding is a combination of word embeddings and position embeddings. We pre-processed document images with an off-the-shelf OCR toolkit to obtain textual content and corresponding 2D position information. We initialize *the word embeddings* with a word embedding matrix from a pre-trained model RoBERTa [36]. *The position embeddings* include 1D position and 2D layout position embeddings, where the 1D position refers to the index of tokens within the text sequence, and the **2D layout position** refers to the bounding box coordinates of the text

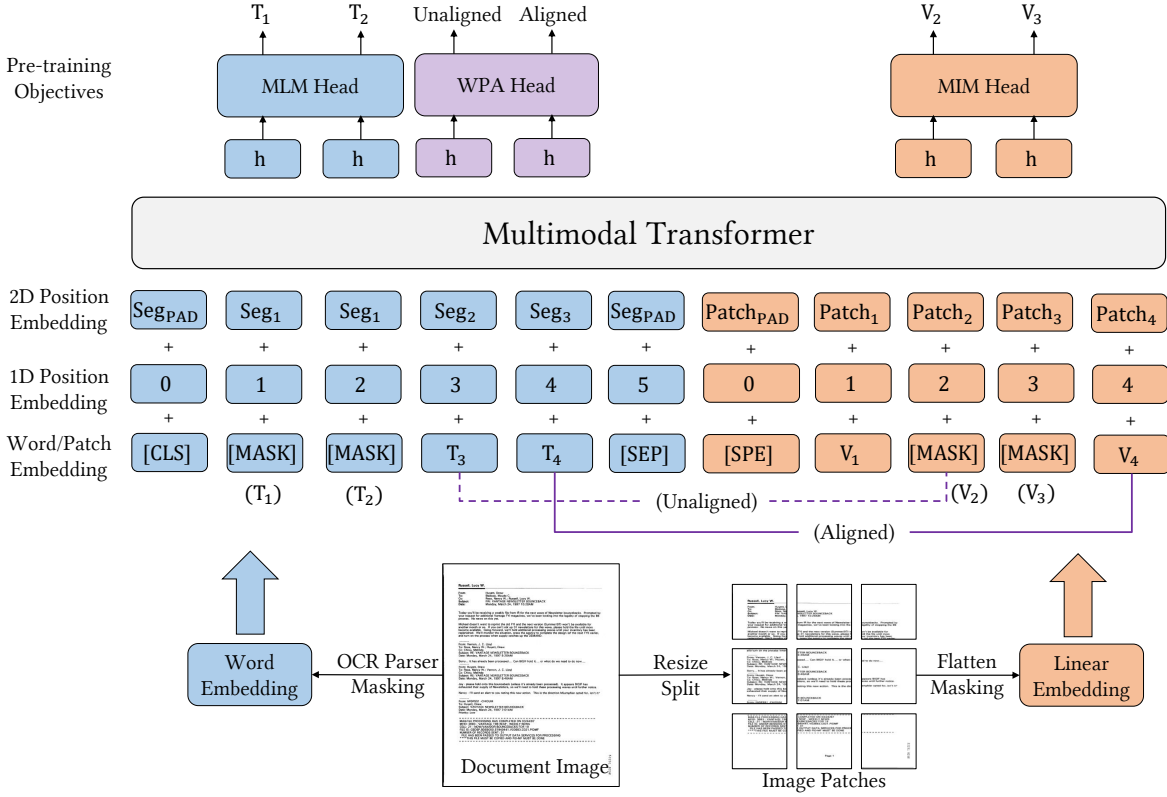


Figure 3: The architecture and pre-training objectives of LayoutLMv3. LayoutLMv3 is a pre-trained multimodal Transformer for Document AI with unified text and image masking objectives. Given an input document image and its corresponding text and layout position information, the model takes the linear projection of patches and word tokens as inputs and encodes them into contextualized vector representations. LayoutLMv3 is pre-trained with discrete token reconstructive objectives of Masked Language Modeling (MLM) and Masked Image Modeling (MIM). Additionally, LayoutLMv3 is pre-trained with a Word-Patch Alignment (WPA) objective to learn cross-modal alignment by predicting whether the corresponding image patch of a text word is masked. “Seg” denotes segment-level positions. “[CLS]”, “[MASK]”, “[SEP]” and “[SPE]” are special tokens.

sequence. Following the LayoutLM, we normalize all coordinates by the size of images, and use embedding layers to embed x-axis, y-axis, width and height features separately [54]. The LayoutLM and LayoutLMv2 adopt word-level layout positions, where each word has its positions. Instead, we adopt segment-level layout positions that words in a segment share the same 2D position since the words usually express the same semantic meaning [28].

Image Embedding. Existing multimodal models in Document AI either extract CNN grid features [2, 56] or rely on an object detector like Faster R-CNN [44] to extract region features [14, 31, 40, 54] for image embeddings, which accounts for heavy computation bottleneck or require region supervision. Inspired by ViT [11] and ViLT [22], we represent document images with linear projection features of image patches before feeding them into the multimodal Transformer. Specifically, we resize a document image into $H \times W$ and denote the image with $I \in \mathbb{R}^{C \times H \times W}$, where C , H and W are the channel size, width and height of the image respectively. We then split the image into a sequence of uniform $P \times P$ patches, linearly project the image patches to D dimensions and flatten them into a sequence of vectors, which length is $M = HW/P^2$. Then we add

learnable 1D position embeddings to each patch since we have not observed improvements from using 2D position embeddings in our preliminary experiments. LayoutLMv3 is the first multimodal model in Document AI that does not rely on CNNs to extract image features, which is vital to Document AI models to reduce parameters or remove complex pre-processing steps.

We insert semantic 1D relative position and spatial 2D relative position as bias terms in self-attention networks for text and image modalities following LayoutLMv2[56].

2.2 Pre-training Objectives

LayoutLMv3 is pre-trained with the MLM, MIM, and WPA objectives to learn multimodal representation in a self-supervised learning manner. Full pre-training objectives of LayoutLMv3 is defined as $L = L_{MLM} + L_{MIM} + L_{WPA}$.

Objective I: Masked Language Modeling (MLM). For the language side, our MLM is inspired by the masked language modeling in BERT [9] and masked visual-language modeling in LayoutLM [54] and LayoutLMv2 [56]. We mask 30% of text tokens with

a span masking strategy with span lengths drawn from a Poisson distribution ($\lambda = 3$) [21, 27]. The pre-training objective is to maximize the log-likelihood of the correct masked text tokens y_ℓ based on the contextual representations of corrupted sequences of image tokens $\mathbf{X}^{M'}$ and text tokens $\mathbf{Y}^{L'}$, where M' and L' represent the masked positions. We denote parameters of the Transformer model with θ and minimize the subsequent cross-entropy loss:

$$L_{MLM}(\theta) = - \sum_{\ell=1}^{L'} \log p_\theta(y_\ell | \mathbf{X}^{M'}, \mathbf{Y}^{L'}) \quad (1)$$

As we keep the layout information unchanged, this objective facilitates the model to learn the correspondence between layout information and text and image context.

Objective II: Masked Image Modeling (MIM). To encourage the model to interpret visual content from contextual text and image representations, we adapt the MIM pre-training objective in BEiT [3] to our multimodal Transformer model. The MIM objective is a symmetry to the MLM objective, that we randomly mask a percentage of about 40% image tokens with the blockwise masking strategy [3]. The MIM objective is driven by a cross-entropy loss to reconstruct the masked image tokens \mathbf{x}_m under the context of their surrounding text and image tokens.

$$L_{MIM}(\theta) = - \sum_{m=1}^{M'} \log p_\theta(\mathbf{x}_m | \mathbf{X}^{M'}, \mathbf{Y}^{L'}) \quad (2)$$

The labels of image tokens come from an image tokenizer, which can transform dense image pixels into discrete tokens according to a visual vocabulary [43]. Thus MIM facilitates learning high-level layout structures rather than noisy low-level details.

Objective III: Word-Patch Alignment (WPA). For documents, each text word corresponds to an image patch. As we randomly mask text and image tokens with MLM and MIM respectively, there is no explicit alignment learning between text and image modalities. We thus propose a WPA objective to learn a fine-grained alignment between text words and image patches. The WPA objective is to predict whether the corresponding image patches of a text word are masked. Specifically, we assign an *aligned* label to an *unmasked* text token when its corresponding image tokens are also unmasked. Otherwise, we assign an *unaligned* label. We exclude the *masked* text tokens when calculating WPA loss to prevent the model from learning a correspondence between masked text words and image patches. We use a two-layer MLP head that inputs contextual text and image and outputs the binary aligned/unaligned labels with a binary cross-entropy loss:

$$L_{WPA}(\theta) = - \sum_{\ell=1}^{L-L'} \log p_\theta(z_\ell | \mathbf{X}^{M'}, \mathbf{Y}^{L'}), \quad (3)$$

where $L - L'$ is the number of unmasked text tokens, z_ℓ is the binary label of language token in the ℓ position.

3 EXPERIMENTS

3.1 Model Configurations

The network architecture of LayoutLMv3 follows that of LayoutLM [54] and LayoutLMv2 [56] for a fair comparison. We use base and large model sizes for LayoutLMv3. LayoutLMv3_{BASE} adopts a 12-layer

Transformer encoder with 12-head self-attention, hidden size of $D = 768$, and 3,072 intermediate size of feed-forward networks. LayoutLMv3_{LARGE} adopts a 24-layer Transformer encoder with 16-head self-attention, hidden size of $D = 1,024$, and 4,096 intermediate size of feed-forward networks. To pre-process the text input, we tokenize the text sequence with Byte-Pair Encoding (BPE) [46] with a maximum sequence length $L = 512$. We add a [CLS] and a [SEP] token at the beginning and end of each text sequence. When the length of the text sequence is shorter than L , we append [PAD] tokens to it. The bounding box coordinates of these special tokens are all zeros. The parameters for image embedding are $C \times H \times W = 3 \times 224 \times 224$, $P = 16$, $M = 196$.

We adopt distributed and mixed-precision training to reduce memory costs and speed up training procedures. We also use a gradient accumulation mechanism to split the batch of samples into several mini-batches to overcome memory constraints for large batch sizes. We further use a gradient checkpointing technique for document layout analysis to reduce memory costs. To stabilize training, we follow CogView [10] to change the computation of attention to $\text{softmax}\left(\frac{\mathbf{Q}^T \mathbf{K}}{\sqrt{d}}\right) = \text{softmax}\left(\left(\frac{\mathbf{Q}^T}{\alpha \sqrt{d}} \mathbf{K} - \max\left(\frac{\mathbf{Q}^T}{\alpha \sqrt{d}} \mathbf{K}\right)\right) \times \alpha\right)$, where α is 32.

3.2 Pre-training LayoutLMv3

To learn a universal representation for various document tasks, we pre-train LayoutLMv3 on a large IIT-CDIP dataset. The **IIT-CDIP Test Collection 1.0** is a large-scale scanned document image dataset, which contains about 11 million document images and can split into 42 million pages [26]. We only use 11 million of them to train LayoutLMv3. We do not do image augmentation following LayoutLM models [54, 56]. For the multimodal Transformer encoder along with the text embedding layer, LayoutLMv3 is initialized from the pre-trained weights of RoBERTa [36]. Our image tokenizer is initialized from a pre-trained image tokenizer in DiT, a self-supervised pre-trained document image Transformer model [30]. The vocabulary size of image tokens is 8,192. We randomly initialized the rest model parameters. We pre-train LayoutLMv3 using Adam optimizer [23] with a batch size of 2,048 for 500,000 steps. We use a weight decay of $1e - 2$, and $(\beta_1, \beta_2) = (0.9, 0.98)$. For the LayoutLMv3_{BASE} model, we use a learning rate of $1e - 4$, and we linearly warm up the learning rate over the first 4.8% steps. For LayoutLMv3_{LARGE}, the learning rate and warm-up ratio are $5e - 5$ and 10%, respectively.

3.3 Fine-tuning on Multimodal Tasks

We compare LayoutLMv3 with typical self-supervised pre-training approaches and categorize them by their pre-training modalities.

- **[T] text modality:** BERT [9] and RoBERTa [36] are typical pre-trained language models which only use text information with Transformer architecture. We use FUNSD and RVL-CDIP results of the RoBERTa from LayoutLM [54] and results of BERT from LayoutLMv2 [56]. We reproduce and report the CORD and DocVQA results of the RoBERTa.
- **[T+L] text and layout modalities:** LayoutLM incorporates layout information by adding word-level spatial embeddings to embeddings of BERT [54]. StructuralLM leverages segment-level layout information [28]. BROS encodes relative layout

Table 1: Comparison with existing published models on the CORD [39], FUNSD [20], RVL-CDIP [16], and DocVQA [38] datasets. “T/L/I” denotes “text/layout/image” modality. “R/G/P” denotes “region/grid/patch” image embedding. We multiply all values by a hundred for better readability. [†]In the UDoc paper [14], the CORD splits are 626/247 receipts for training/test instead of the official 800/100 training/test receipts adopted by other works. Thus the score[†] is not directly comparable to other scores. Models denoted with [‡] use more data to train DocVQA and are expected to score higher. For example, TILT introduces one more supervised training stage on more QA datasets [40]. StructuralLM additionally uses the validation set in training [28].

| Model | Parameters | Modality | Image Embedding | FUNSD F1↑ | CORD F1↑ | RVL-CDIP Accuracy↑ | DocVQA ANLS↑ |
|--|------------|-----------|------------------------|--------------|--------------------|-----------------------|--------------------|
| BERT _{BASE} [9] | 110M | T | None | 60.26 | 89.68 | 89.81 | 63.72 |
| RoBERTa _{BASE} [36] | 125M | T | None | 66.48 | 93.54 | 90.06 | 66.42 |
| BROS _{BASE} [17] | 110M | T+L | None | 83.05 | 95.73 | - | - |
| LiLT _{BASE} [50] | - | T+L | None | 88.41 | 96.07 | 95.68* | - |
| LayoutLM _{BASE} [54] | 160M | T+L+I (R) | ResNet-101 (fine-tune) | 79.27 | - | 94.42 | - |
| SelfDoc [31] | - | T+L+I (R) | ResNeXt-101 | 83.36 | - | 92.81 | - |
| UDoc [14] | 272M | T+L+I (R) | ResNet-50 | 87.93 | 98.94 [†] | 95.05 | - |
| TILT _{BASE} [40] | 230M | T+L+I (R) | U-Net | - | 95.11 | 95.25 | 83.92 [‡] |
| XYLayoutLM _{BASE} [15] | - | T+L+I (G) | ResNeXt-101 | 83.35 | - | - | - |
| LayoutLMv2 _{BASE} [56] | 200M | T+L+I (G) | ResNeXt101-FPN | 82.76 | 94.95 | 95.25 | 78.08 |
| DocFormer _{BASE} [2] | 183M | T+L+I (G) | ResNet-50 | 83.34 | 96.33 | 96.17 | - |
| LayoutLMv3_{BASE} (Ours) | 133M | T+L+I (P) | Linear | 90.29 | 96.56 | 95.44 | 78.76 |
| BERT _{LARGE} [9] | 340M | T | None | 65.63 | 90.25 | 89.92 | 67.45 |
| RoBERTa _{LARGE} [36] | 355M | T | None | 70.72 | 93.80 | 90.11 | 69.52 |
| LayoutLM _{LARGE} [54] | 343M | T+L | None | 77.89 | - | 91.90 | - |
| BROS _{LARGE} [17] | 340M | T+L | None | 84.52 | 97.40 | - | - |
| StructuralLM _{LARGE} [28] | 355M | T+L | None | 85.14 | - | 96.08 | 83.94 [‡] |
| FormNet [25] | 217M | T+L | None | 84.69 | - | - | - |
| FormNet [25] | 345M | T+L | None | - | 97.28 | - | - |
| TILT _{LARGE} [40] | 780M | T+L+I (R) | U-Net | - | 96.33 | 95.52 | 87.05 [‡] |
| LayoutLMv2 _{LARGE} [56] | 426M | T+L+I (G) | ResNeXt101-FPN | 84.20 | 96.01 | 95.64 | 83.48 |
| DocFormer _{LARGE} [2] | 536M | T+L+I (G) | ResNet-50 | 84.55 | 96.99 | 95.50 | - |
| LayoutLMv3_{LARGE} (Ours) | 368M | T+L+I (P) | Linear | 92.08 | 97.46 | 95.93 | 83.37 |

* LiLT uses image features with ResNeXt101-FPN backbone in fine-tuning RVL-CDIP.

positions [17]. LiLT fine-tunes on different languages with pre-trained textual models [50]. FormNet leverages the spatial relationship between tokens in a form [25].

- **[T+L+I (R)] text, layout and image modalities with Faster R-CNN region features:** This line of works extract image region features from RoI heads in the Faster R-CNN model [44]. Among them, LayoutLM [54] and TILT [40] use OCR words’ bounding box to serve as region proposals and add the region features to corresponding text embeddings. SelfDoc [31] and UDoc [14] use document object proposals and concatenate region features with text embeddings.
- **[T+L+I (G)] text, layout and image modalities with CNN grid features:** LayoutLMv2 [56] and DocFormer [2] extract image grid features with a CNN backbone without object detection. XYLayoutLM [15] adopts the architecture of LayoutLMv2 and improves layout representation.
- **[T+L+I (P)] text, layout, and image modalities with linear patch features:** LayoutLMv3 replaces CNN backbones with simple linear embedding to encode image patches.

We fine-tune LayoutLMv3 on multimodal tasks on publicly available benchmarks. Results are shown in Table 1.

Task 1: Form and Receipt Understanding. Form and receipt understanding tasks require extracting and structuring forms and receipts’ textual content. The tasks are a sequence labeling problem aiming to tag each word with a label. **We predict the label of the last hidden state of each text token with a linear layer and an MLP classifier for form and receipt understanding tasks, respectively.**

We conduct experiments on the FUNSD dataset and the CORD dataset. The **FUNSD** [20] is a noisy scanned form understanding dataset sampled from the RVL-CDIP dataset [16]. The FUNSD dataset contains 199 documents with comprehensive annotations for 9,707 semantic entities. We focus on the semantic entity labeling task on the FUNSD dataset to assign each semantic entity a label among “question”, “answer”, “header” or “other”. The training and test splits contain 149 and 50 samples, respectively. **CORD** [39] is a receipt key information extraction dataset with 30 semantic labels defined under 4 categories. It contains 1,000 receipts of 800 training, 100 validation, and 100 test examples. We use officially-provided images and OCR annotations. We fine-tune LayoutLMv3 for 1,000

Table 2: Document layout analysis mAP @ IOU [0.50:0.95] on PubLayNet validation set. All models use only information from the vision modality. LayoutLMv3 outperforms the compared ResNets [14, 59] and vision Transformer [30] backbones.

| Model | Framework | Backbone | Text | Title | List | Table | Figure | Overall |
|---|---------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| PubLayNet[59] | Mask R-CNN | ResNet-101 | 91.6 | 84.0 | 88.6 | 96.0 | 94.9 | 91.0 |
| DiT _{BASE} [30] | Mask R-CNN | Transformer | 93.4 | 87.1 | 92.9 | 97.3 | 96.7 | 93.5 |
| UDoc [14] | Faster R-CNN | ResNet-50 | 93.9 | 88.5 | 93.7 | 97.3 | 96.4 | 93.9 |
| DiT _{BASE} [30] | Cascade R-CNN | Transformer | 94.4 | 88.9 | 94.8 | 97.6 | 96.9 | 94.5 |
| LayoutLMv3_{BASE} (Ours) | Cascade R-CNN | Transformer | 94.5 | 90.6 | 95.5 | 97.9 | 97.0 | 95.1 |

steps with a learning rate of $1e-5$ and a batch size of 16 for FUNSD, and $5e-5$ and 64 for CORD.

We report F1 scores for this task. For the large model size, the LayoutLMv3 achieves an F1 score of 92.08 on the FUNSD dataset, which significantly outperforms the SOTA result of 85.14 provided by StructuralLM [28]. Note that LayoutLMv3 and StructuralLM use segment-level layout positions, while the other works use word-level layout positions. Using segment-level positions may benefit the semantic entity labeling task on FUNSD [28], so the two types of work are not directly comparable. The LayoutLMv3 also achieves SOTA F1 scores on the CORD dataset for both base and large model sizes. The results show that LayoutLMv3 can significantly benefit the text-centric form and receipt understanding tasks.

Task II: Document Image Classification. The document image classification task aims to predict the category of document images. We feed the output hidden state of the special classification token ([CLS]) into an MLP classifier to predict the class labels.

We conduct experiments on the **RVL-CDIP** dataset. It is a subset of the IIT-CDIP collection labeled with 16 categories [16]. RVL-CDIP dataset contains 400,000 document images, among them 320,000 are training images, 40,000 are validation images, and 40,000 are test images. We extract text and layout information using Microsoft Read API. We fine-tune LayoutLMv3 for 20,000 steps with a batch size of 64 and a learning rate of $2e-5$.

The evaluation metric is the overall classification accuracy. LayoutLMv3 achieves better or comparable results with a much smaller model size than previous works. For example, compared to LayoutLMv2, LayoutLMv3 achieves an absolute improvement of 0.19% and 0.29% in the base model and large model size, respectively, with a much simpler image embedding (i.e., Linear vs. ResNeXt101-FPN). The results show that our simple image embeddings can achieve desirable results on image-centric tasks.

Task III: Document Visual Question Answering. Document visual question answering requires a model to take a document image and a question as input and output an answer [38]. We formalize this task as an extractive QA problem, where the model predicts start and end positions by classifying the last hidden state of each text token with a binary classifier.

We conduct experiments on the **DocVQA** dataset, a standard dataset for visual question answering on document images [38]. The official partition of the DocVQA dataset consists of 10,194/1,286/1,287 images and 39,463/5,349/5,188 questions for training/validation/test set, respectively. We train our model on the training set, evaluate the model on the test set, and report results by submitting them to the official evaluation website. We use Microsoft Read API

to extract text and bounding boxes from images and use heuristics to find given answers in the extracted text as in LayoutLMv2. We fine-tune LayoutLMv3_{BASE} for 100,000 steps with a batch size of 128, a learning rate of $3e-5$, and a warmup ratio of 0.048. For LayoutLMv3_{LARGE}, the step size, batch size, learning rate and warmup ratio are 200,000, 32, $1e-5$, and 0.1, respectively.

We report the commonly-used edit distance-based metric ANLS (also known as Average Normalized Levenshtein Similarity). The LayoutLMv3_{BASE} improves the ANLS score of LayoutLMv2_{BASE} from 78.08 to 78.76, with much simpler image embedding (i.e., from ResNeXt101-FPN to Linear embedding). The LayoutLMv3_{LARGE} further gains an absolute ANLS score of 4.61 over LayoutLMv3_{BASE}. The results show that LayoutLMv3 is effective for the document visual question answering task.

3.4 Fine-tuning on a Vision Task

To demonstrate the generality of LayoutLMv3 from the multimodal domain to the visual domain, we transfer LayoutLMv3 to a **document layout analysis** task. This task is about detecting the layouts of unstructured digital documents by providing bounding boxes and categories such as tables, figures, texts, etc. This task helps parse the documents into a machine-readable format for downstream applications. We model this task as an object detection problem without text embedding, which is effective in existing works [14, 30, 59]. We integrate the LayoutLMv3 as feature backbone in the Cascade R-CNN detector [4] with FPN [34] implemented using the Detectron2 [53]. We adopt the standard practice to extract single-scale features from different Transformer layers, such as layers 4, 6, 8, and 12 of the LayoutLMv3 base model. We use resolution-modifying modules to convert the single-scale features into the multiscale FPN features [1, 30, 33].

We conduct experiments on **PubLayNet** dataset [59]. The dataset contains research paper images annotated with bounding boxes and polygonal segmentation across five document layout categories: text, title, list, figure, and table. The official splits contain 335,703 training images, 11,245 validation images, and 11,405 test images. We train our model on the training split and evaluate our model on the validation split following standard practice [14, 30, 59]. We train our model for 60,000 steps using the AdamW optimizer with 1,000 warm-up steps and a weight decay of 0.05 following DiT [30]. Since LayoutLMv3 is pre-trained with inputs from both vision and language modalities, we use a larger batch size of 32 and a lower learning rate of $2e-4$ empirically. We do not use flipping or cropping augmentation strategy in the fine-tuning stage to be consistent

Table 3: Ablation study on image embeddings and pre-training objectives on typical text-centric tasks (form and receipt understanding on FUNSD and CORD) and image-centric tasks (document image classification on RVL-CDIP and document layout analysis on PubLayNet). All models were trained at BASE size on 1 million data for 150,000 steps with learning rate $3e-4$.

| # | Image Embed | Parameters | Pre-training Objective(s) | FUNSD F1↑ | CORD F1↑ | RVL-CDIP Accuracy↑ | PubLayNet MAP↑ |
|---|-------------|------------|---------------------------|--------------|--------------|--------------------|-----------------|
| 1 | None | 125M | MLM | 88.64 | 96.27 | 95.33 | Not Applicable |
| 2 | Linear | 126M | MLM | 89.39 | 96.11 | 95.00 | Loss Divergence |
| 3 | Linear | 132M | MLM+MIM | 89.19 | 96.30 | 95.42 | 94.38 |
| 4 | Linear | 133M | MLM+MIM+WPA | 89.78 | 96.49 | 95.53 | 94.43 |

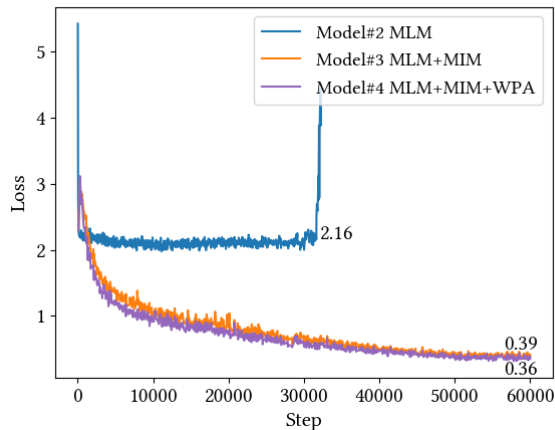


Figure 4: Loss convergence curves of fine-tuning the ablated models of LayoutLMv3 on PubLayNet dataset. The loss of model #2 did not converge. By incorporating the MIM objective, the loss converges normally. The WPA objective further decreases the loss. Best viewed in color.

with our pre-training stage. We do not use relative positions in self-attention networks as DiT.

We measure the performance using the mean average precision (MAP) @ intersection over union (IOU) [0.50:0.95] of bounding boxes and report results in Table 2. We compare with the ResNets [14, 59] and the concurrent vision Transformer [30] backbones. LayoutLMv3 outperforms the other models in all metrics, achieving an overall mAP score of 95.1. LayoutLMv3 achieves a high gain in the “Title” category. Since titles are typically much smaller than other categories and can be identified by their textual content, we attribute this improvement to our incorporation of language modality in pre-training LayoutLMv3. These results demonstrate the generality and superiority of LayoutLMv3.

3.5 Ablation Study

In Table 3 we study the effect of our image embeddings and pre-training objectives. We first build a baseline model #1 that uses text and layout information, pre-trained with MLM objective. Then we use linearly projected image patches as the image embedding of the baseline model, denoted as model #2. We further pre-train model

#2 with MIM and WPA objectives step by step and denote the new models as #3 and #4, respectively.

In Figure 4, we visualize losses of models #2, #3, and #4 when fine-tuned on the PubLayNet dataset with a batch size of 16 and a learning rate of $2e-4$. We have tried to train the model #2 with learning rates of $\{1e-4, 2e-4, 4e-4\}$ combined with batch sizes of $\{16, 32\}$, but the loss of model #2 did not converge and the mAP score on PubLayNet is near zero.

Effect of Linear Image Embedding. We observe that model #1 without image embedding has achieved good results on some tasks. This suggests that language modality, including text and layout information, plays a vital role in document understanding. However, the results are still unsatisfactory. Moreover, model #1 cannot conduct some image-centric document analysis tasks without vision modality. For example, the vision modality is critical for the document layout analysis task on PubLayNet because bounding boxes are tightly integrated with images. Our simple design of linear image embedding combined with appropriate pre-training objectives can consistently improve not only image-centric tasks, but also some text-centric tasks further.

Effect of MIM pre-training objective. Simply concatenating linear image embedding with text embedding as input to model #2 deteriorates performance on CORD and RVL-CDIP, while the loss on PubLayNet diverges. We speculate that the model failed to learn meaningful visual representation on the linear patch embeddings without any pre-training objective associated with image modality. The MIM objective mitigates this problem by preserving the image information until the last layer of the model by randomly masking out a portion of input image patches and reconstructing them in the output [22]. Comparing the results of model #3 and model #2, the MIM objective benefits CORD and RVL-CDIP. As simply using linear image embedding has improved FUNSD, MIM does not further contribute to FUNSD. By incorporating the MIM objective in training, the loss converges when fine-tuning PubLayNet as shown in Figure 4, and we obtain a desirable mAP score. The results indicate that MIM can help regularize the training. Thus MIM is critical for vision tasks like document layout analysis on PubLayNet.

Effect of WPA pre-training objective. By comparing models #3 and #4 in Table 3, we observe that the WPA objective consistently improves all tasks. Moreover, the WPA objective decreases the loss of the vision task on PubLayNet in Figure 4. These results confirm the effectiveness of WPA not only in cross-modal representation learning, but also in image representation learning.

Parameter Comparisons. The table shows that incorporating image embedding for a 16×16 patch projection ($\#1 \rightarrow \#2$) introduces only 0.6M parameters. The parameters are negligible compared to the parameters of CNN backbones (e.g., 44M for ResNet-101). A MIM head and a WPA head introduce 6.9M and 0.6M parameters in the pre-training stage. The parameter overhead introduced by image embedding is marginal compared to the MLM head, which has 39.2M parameters for a text vocabulary size of 50,265. We did not take count of the image tokenizer when calculating parameters as the tokenizer is a standalone module for generating the labels of MIM but is not integrated into the Transformer backbone.

4 RELATED WORK

Multimodal self-supervised pre-training technique has made a rapid progress in *document intelligence* due to its successful applications of document layout and image representation learning [2, 13–15, 17, 25, 28, 31, 32, 40, 41, 50, 52, 54–56]. LayoutLM and following works joint layout representation learning by encoding spatial coordinates of text [17, 25, 28, 54]. Various works then joint image representation learning by combining CNNs with Transformer [49] self-attention networks. These works either extract CNN grid features [2, 56] or rely on an object detector to extract region features [14, 31, 40, 54], which accounts for heavy computation bottleneck or requires region supervision. In the field of *natural images vision-and-language pre-training* (VLP), research works have seen a shift from region features [5, 47, 48] to grid features [19] to lift limitations of pre-defined object classes and region supervision. Inspired by vision Transformer (ViT) [11], there have also been recent efforts in VLP without CNNs to overcome the weakness of CNN. Still, most rely on separate self-attention networks to learn visual features; thus, their computational cost is not reduced [12, 29, 57]. An exception is ViLT, which learns visual features with a lightweight linear layer and significantly cuts down the model size and running time [22]. Inspired by ViLT, our LayoutLMv3 is the first multimodal model in Document AI that utilizes image embeddings without CNNs.

Reconstructive pre-training objectives revolutionized representation learning. In *NLP* research, BERT firstly proposed “masked language modeling” (MLM) to learn bidirectional representations and advanced the state of the arts on broad language understanding tasks [9]. In the field of CV, Masked Image Modeling (MIM) aims to learn rich visual representations via predicting masked content conditioning in visible context. For example, ViT reconstructs the mean color of masked patches, which leads to performance gains in ImageNet classification [11]. BEiT reconstructs visual tokens learned by a discrete VAE, achieving competitive results in image classification and semantic segmentation [3]. DiT extends BEiT to document images to document layout analysis [30].

Inspired by MLM and MIM, researchers in the field of *vision-and-language* have explored **reconstructive objectives for multimodal representation learning**. Whereas most well-performing vision-and-language pre-training (VLP) models use the MLM proposed by BERT on text modality, they differ in their pre-training objectives for the image modality. There are three variants of MIM corresponding to different image embeddings: masked region modeling (MRM), masked grid modeling (MGM), and masked patch modeling

(MPM). MRM has been proven to be effective in regressing original region features [5, 31, 48] or classifying object labels [5, 37, 48] for masked regions. MGM has also been explored in the SOHO, whose objective is to predict the mapping index in a visual dictionary for masked grid features [19]. For patch-level image embedding, Visual Parsing [57] proposed to mask visual tokens according to the attention weights in their self-attention image encoder, which does not apply to simple linear image encoders. ViLT [22] and METER [12] attempt to leverage MPM similar to ViT [11] and BEiT [3], which respectively reconstruct the mean color and discrete tokens in visual vocabularies for image patches, but resulted in degraded performance on downstream tasks. Our LayoutLMv3 firstly demonstrates the effectiveness of MIM for linear patch image embedding.

Various **cross-modal objectives** are further developed for vision and language (VL) alignment learning in multimodal models. Image-text matching is widely used to learn a coarse-grained VL alignment [2, 5, 19, 22, 56]. To learn a fine-grained VL alignment, UNITER proposes a word-region alignment objective based on optimal transports, which calculates the minimum cost of transporting the contextualized image embeddings to word embeddings [5]. ViLT extends this objective to patch-level image embeddings [22]. Unlike natural images, document images imply an explicit fine-grained alignment relationship between text words and image areas. Using this relationship, UDoc uses contrastive learning and similarity distillation to align the image and text belonging to the same area [14]. LayoutLMv2 covers some text lines in raw images and predicts whether each text token is covered [56]. In contrast, we naturally utilize the masking operations in MIM to construct aligned/unaligned pairs in an effective and unified way.

5 CONCLUSION AND FUTURE WORK

In this paper, we present LayoutLMv3 to pre-train the multimodal Transformer for Document AI, which redesigns the model architecture and pre-training objectives for LayoutLM. Distinguishing from the existing multimodal model in Document AI, LayoutLMv3 does not rely on a pre-trained CNN or Faster R-CNN backbone to extract visual features, significantly saving parameters and eliminating region annotations. We use unified text and image masking pre-training objectives: masked language modeling, masked image modeling, and word-patch alignment, to learn multimodal representations. Extensive experimental results have demonstrated the generality and superiority of LayoutLMv3 for both text-centric and image-centric Document AI tasks with the simple architecture and unified objectives. In future research, we will investigate scaling up pre-trained models so that the models can leverage more training data to drive SOTA results further. In addition, we will explore few-shot and zero-shot learning capabilities to facilitate more real-world business scenarios in the Document AI industry.

6 ACKNOWLEDGEMENT

We are grateful to Yiheng Xu for fruitful discussions and inspiration. This work was supported by the NSFC (U1811461) and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant NO.2016ZT06D211.

Table 4: Visual information extraction in Chinese F1 score on the EPHOIE test set.

| Model | Subject | Test Time | Name | School | #Examination | #Seat | Class | #Student | Grade | Score | Mean |
|---|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| BILSTM+CRF [24] | 98.51 | 100.0 | 98.87 | 98.80 | 75.86 | 72.73 | 94.04 | 84.44 | 98.18 | 69.57 | 89.10 |
| GCN-based [35] | 98.18 | 100.0 | 99.52 | 100.0 | 88.17 | 86.00 | 97.39 | 80.00 | 94.44 | 81.82 | 92.55 |
| GraphIE [42] | 94.00 | 100.0 | 95.84 | 97.06 | 82.19 | 84.44 | 93.07 | 85.33 | 94.44 | 76.19 | 90.26 |
| TRIE [58] | 98.79 | 100.0 | 99.46 | 99.64 | 88.64 | 85.92 | 97.94 | 84.32 | 97.02 | 80.39 | 93.21 |
| VIES [51] | 99.39 | 100.0 | 99.67 | 99.28 | 91.81 | 88.73 | 99.29 | 89.47 | 98.35 | 86.27 | 95.23 |
| StrucTexT [32] | 99.25 | 100.0 | 99.47 | 99.83 | 97.98 | 95.43 | 98.29 | 97.33 | 99.25 | 93.73 | 97.95 |
| LayoutLMv3-Chinese_{BASE} (Ours) | 98.99 | 100.0 | 99.77 | 99.20 | 100.0 | 100.0 | 98.82 | 99.78 | 98.31 | 97.27 | 99.21 |

A APPENDIX

A.1 LayoutLMv3 in Chinese

Pre-training LayoutLMv3 in Chinese. To demonstrate the effectiveness of LayoutLMv3 in not only English but also in the Chinese language, we pre-train a LayoutLMv3-Chinese model in base size. It is trained on 50 million document pages in Chinese. We collect large-scale Chinese documents by downloading publicly available digital-born documents and following the principles of Common Crawl (<https://commoncrawl.org/>) to process these documents. For the multimodal Transformer encoder along with the text embedding layer, LayoutLMv3-Chinese is initialized from the pre-trained weights of XLM-R [7]. We randomly initialized the rest model parameters. Other training setting is the same as LayoutLMv3.

Fine-tuning on Visual Information Extraction. The visual information extraction (VIE) requires extracting key information from document images. The task is a sequence labeling problem aiming to tag each word with a pre-defined label. We predict the label of the last hidden state of each text token with a linear layer.

We conduct experiments on the EPHOIE dataset. The EPHOIE [51] is a visual information extraction dataset consisting of examination paper heads with diverse layouts and backgrounds. It contains 1,494 images with comprehensive annotations for 15,771 Chinese text instances. We focus on a token-level entity labeling task on the EPHOIE dataset to assign each character a label among ten pre-defined categories. The training and test sets contain 1,183 and 311 images, respectively. We fine-tune LayoutLMv3-Chinese for 100 epochs. The batch size is 16, and the learning rate is $5e-5$ with linear warmup over the first epoch.

We report F1 scores for this task and report results in Table 4. The LayoutLMv3-Chinese shows superior performance on most metrics and achieves a SOTA mean F1 score of 99.21%. The results show that LayoutLMv3 significantly benefits the VIE task in Chinese.

REFERENCES

- [1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. 2021. Xcit: Cross-covariance image transformers. In *NeurIPS*.
- [2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. DocFormer: End-to-End Transformer for Document Understanding. In *ICCV*.
- [3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEiT: BERT Pre-Training of Image Transformers. In *ICLR*.
- [4] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *CVPR*.
- [5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.
- [6] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. 2020. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In *EMNLP*.
- [7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Mylé Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*.
- [8] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document AI: Benchmarks, Models and Applications. *arXiv preprint arXiv:2111.08609* (2021).
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.
- [10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.
- [12] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. 2021. An Empirical Study of Training End-to-End Vision-and-Language Transformers. *arXiv preprint arXiv:2111.02387* (2021).
- [13] Łukasz Garmarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. LAMBERT: Layout-Aware Language Modeling for Information Extraction. In *ICDAR*.
- [14] Jiuxiang Gu, Jason Kuen, Vlad Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. UniDoc: Unified Pretraining Framework for Document Understanding. In *NeurIPS*.
- [15] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022. XYLayoutLM: Towards Layout-Aware Multimodal Networks For Visually-Rich Document Understanding. In *CVPR*.
- [16] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval. In *ICDAR*.
- [17] Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. BROS: A Pre-Trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents. In *AAAI*.
- [18] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. 2021. Unifying multimodal transformer for bi-directional image and text generation. In *ACM Multimedia*.
- [19] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*.
- [20] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDARW*.
- [21] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics* 8 (2020), 64–77.
- [22] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.
- [23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [24] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *NAACL HLT*.
- [25] Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction. In *ACL*.
- [26] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a Test Collection for Complex Document Information Processing. In *SIGIR*.

- [27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*.
- [28] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021. StructuralLM: Structural Pre-training for Form Understanding. In *ACL*.
- [29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.
- [30] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. DiT: Self-supervised Pre-training for Document Image Transformer. *arXiv preprint arXiv:2203.02378* (2022).
- [31] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. SelfDoc: Self-Supervised Document Representation Learning. In *CVPR*.
- [32] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021. StrucText: Structured Text Understanding with Multi-Modal Transformers. In *ACM Multimedia*.
- [33] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. 2021. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429* (2021).
- [34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *CVPR*.
- [35] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph Convolution for Multimodal Information Extraction from Visually Rich Documents. In *NAACL HLT*.
- [36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [37] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.
- [38] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.
- [39] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. In *Document Intelligence Workshop at Neural Information Processing Systems*.
- [40] Rafal Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Palka. 2021. Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. In *ICDAR*.
- [41] 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* (2020).
- [42] Yujie Qian. 2019. *A graph-based framework for information extraction*. Ph.D. Dissertation. Massachusetts Institute of Technology.
- [43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*.
- [44] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *TPAMI* 39, 1137–1149.
- [45] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. 2017. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. In *ICLR*.
- [46] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.
- [47] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.
- [48] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*.
- [49] 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 *NeurIPS*.
- [50] Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. In *ACL*.
- [51] Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021. Towards robust visual information extraction in real world: new dataset and novel solution. In *AAAI*.
- [52] Te-Lin Wu, Cheng Li, Mingyang Zhang, Tao Chen, Spurthi Amba Hombaiah, and Michael Bendersky. 2021. LAMPRET: Layout-Aware Multimodal PreTraining for Document Understanding. *arXiv preprint arXiv:2104.08405* (2021).
- [53] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. <https://github.com/facebookresearch/detectron2>.
- [54] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD*.
- [55] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding. *arXiv preprint arXiv:2104.08836* (2021).
- [56] 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. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *ACL*.
- [57] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. 2021. Probing Inter-modality: Visual Parsing with Self-Attention for Vision-and-Language Pre-training. In *NeurIPS*.
- [58] 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 Multimedia*.
- [59] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: largest dataset ever for document layout analysis. In *ICDAR*.