ERNIE-Layout: Layout-Knowledge Enhanced Multi-modal Pre-training for Document Understanding

Anonymous ACL submission

Abstract

We propose ERNIE-Layout, a knowledge enhanced pre-training approach for visual document understanding, which incorporates layout-knowledge into the pre-training of visual document understanding to learn a better joint multi-modal representation of text, layout and image. Previous works directly model serialized tokens from documents according to a raster-scan order, neglecting the importance of the reading order of documents, leading to sub-optimal performance. We incorporate layout-knowledge from Document-Parser into document pre-training, which is used to rearrange the tokens following an order more consistent with human reading habits. And we propose the Reading Order Prediction (ROP) task to enhance the interactions within segments and correlation between segments and a fine-grained cross-modal alignment pre-training task named Replaced Regions Prediction (RRP). ERNIE-Layout attempts to fuse textual and visual features in a unified Transformer model, which is based on our newly proposed spatial-aware disentangled attention mechanism. ERNIE-Layout achieves superior performance on various document understanding tasks, setting new SOTA for four tasks, including information extraction, document classification, document question answering.

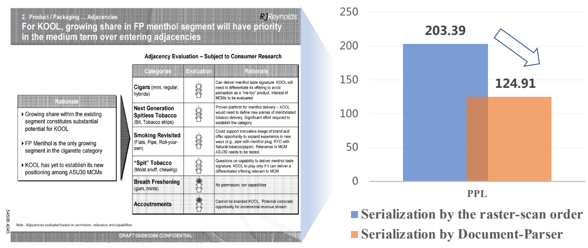


Figure 1: The effect of the knowledge enhanced serialization compared with raster-scan serialization on an example document. Serialized by Document-Parser, the PPL score on the document with complex layout will be significantly reduced. More details are introduced in Section 3.1.

1 Introduction

Visual Document Understanding (VDU) is an important research field that aims to understand various types of digital-born or scanned documents (letter, memo, email, form, invoice, advertisement, etc.) and has attracted great attention from both the industry and the academia due to its various applications. The diversity and the complexity of the formats and layouts in the documents make VDU a more challenging task than the plain-text understanding task.

The early works for VDU (Cheng et al., 2020; Sage et al., 2020; Yang et al., 2016; Katti et al., 2018; Yang et al., 2017; Sarkhel and Nandi, 2019; Palm et al., 2019; Wang et al., 2021) mainly adopt single-modal or shallow multi-modal fusion approaches, which are task-specific and require massive data annotations. Recently, inspired by the development of pre-training techniques in NLP and CV areas, many document pre-training approaches (Xu et al., 2020b,a; Li et al., 2021a,b; Garncarek et al., 2021; Powalski et al., 2021; Appalaraju et al., 2021) have been proposed and shown great improvements for various VDU tasks. As a pioneering work, LayoutLM (Xu et al., 2020b) proposes a document pre-training model which jointly leverages text and layout information, while the visual features from the document image are only utilized during the fine-tuning stage. StructuralLM (Li et al., 2021a) further exploits the segment-level layout instead of the word-level layout. LayoutLMv2 (Xu et al., 2020a) attempts to use the image features during the pre-training stage and adopts a spatial-aware self-attention mechanism and seems to be an improved version of LayoutLM.

However, as an important preprocessing step for all document pre-training methods, the serializing is performed on the OCR results according to a raster-scan order. The raster-scan serialization arranges the tokens by top-left to bottom-right order, which may be inconsistent with human reading habits for documents with complex layouts (multi-column papers, tables, forms, etc.) and leads to sub-optimal performances for the understanding tasks.

Inspired by the pioneering knowledge enhanced pre-training method ERNIE (Sun et al., 2019), in this paper, we present ERNIE-Layout, a layout-knowledge enhanced pre-training approach to improve the performances for document understanding tasks. ERNIE-Layout utilizes serialized input token sequences, which are rearranged by Document-Parser, which is a commercial document layout parser for document analysis. The parser actually provides layout-knowledge, which is the layout analysis of the document. According to this knowledge, the serialized tokens can be rearranged in a more consistent manner with human reading habits. The effect of knowledge enhanced serialization is shown in Figure 1.

We propose the pre-training task Reading Order Prediction (ROP) to enhance the interaction within segments and the correlation between segments, which aims to predict the position of the next token and Replaced Regions Prediction (RRP) to build the fine-grained semantic correspondence between the visual and textual modalities. Furthermore, we integrate a spatial-aware disentangled attention mechanism, inspired by DeBERTa (He et al., 2020), into the encoder-only Transformer, where the attention weights among tokens are computed using disentangled matrices based on their contents, 1D and 2D relative positions.

We conduct experiments on various Visual Document Understanding tasks and find that ERNIE-Layout outperforms previous best approaches on most downstream tasks, proving the effectiveness of our method.

The contributions of this paper are summarized as follows:

- To the best of our knowledge, ERNIE-Layout is the first work that incorporates layout-knowledge to enhance the pre-training for document understanding.

- ERNIE-Layout constructs Reading Order Prediction to enhance the interaction within segments and correlation between segments, and Replaced Regions Prediction to strengthen the alignment between different modalities. ERNIE-Layout adopts our newly proposed spatial-aware disentangled attention mechanism in the Transformer encoder to improve the interaction between semantic features and spatial features.

- ERNIE-Layout achieves state-of-the-art results on various downstream document understanding tasks, including Information Extraction and Document Question Answering.

2 Related Work

Inspired by the success of pre-training techniques in NLP and CV areas, researchers attempt to utilize the pre-training and fine-tuning paradigm for document understanding tasks. Existing visual document pre-training methods contribute their efforts in two aspects: model architecture and pre-training task.

Model Architecture Previous document pre-training models mainly adopt an encoder-only structure (Xu et al., 2020b; Li et al., 2021a; Xu et al., 2020a; Appalaraju et al., 2021; Li et al., 2021b; Garncarek et al., 2021; Powalski et al., 2021), using a Transformer to fuse text, image and layout information. LayoutLM (Xu et al., 2020b) models the interaction between text and layout, while only using image information for downstream tasks. Based on LayoutLM, StructralLM (Li et al., 2021a) leverages segment-level layout instead of word-level. LayoutLMv2 (Xu et al., 2020a) proposes to add image features during the pre-training stage and uses spatial-aware attention, which is an improved version of LayoutLM. DocFormer (Appalaraju et al., 2021) designs a multi-modal attention layer capable of fusing text, vision and spatial features in a document. More recently, TILT (Powalski et al., 2021) proposes an encoder-decoder structure model to generate values not included in the input text explicitly.

Pre-Training Task During the pre-training stage, various types of tasks are proposed to learn the correlation of text, image and layout information. The single-modal pre-training tasks aim to learn text, image or layout representation under multi-modal context. LayoutLM (Xu et al., 2020b) and LayoutLMv2 (Xu et al., 2020a) use the Masked Visual-Language Modeling task to reconstruct the entire sequence with the masked sequence as input, which can make the model learn better text

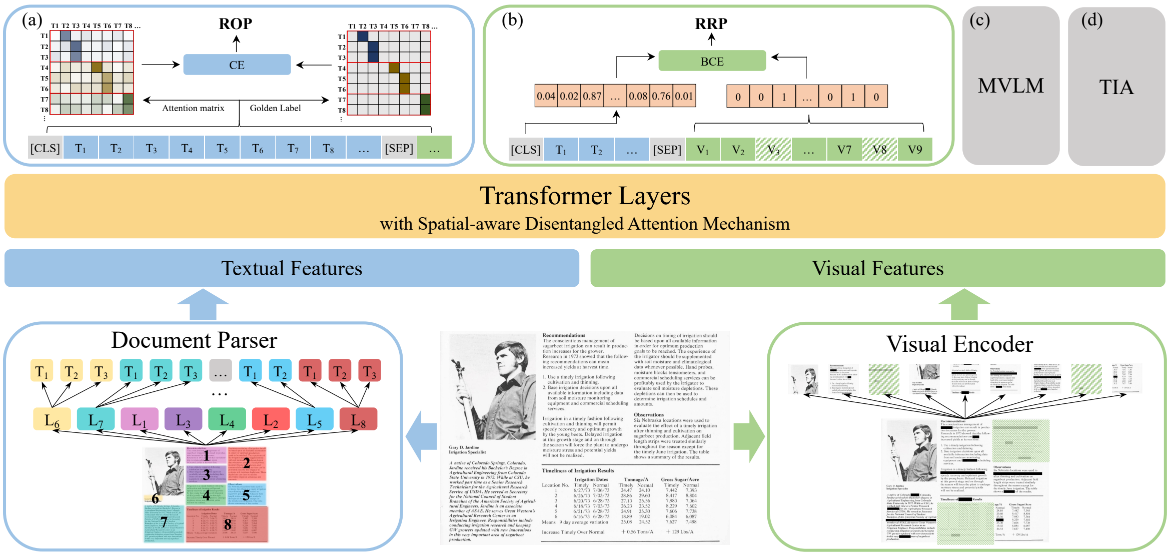


Figure 2: Conceptual overview of ERNIE-Layout. The pre-training tasks consist of: (a) ROP: Reading Order Prediction; (b) RRP: Replaced Regions Prediction; (c) MVLM: Masked Visual-Language Modeling; (d) TIA: Text-Image Alignment.

representation with multi-modal features. Learn to Reconstruct (Appalaraju et al., 2021) aims to reconstruct the image using a shallow decoder in the presence of image and text features. StructuralLM (Li et al., 2021a) proposes the Cell Position Classification task, which predicts where the cells are in documents. The cross-modal pre-training tasks aim to learn the correlation of multi-modalities. Text-Image Matching (Xu et al., 2020a) and Text-Image Alignment (Xu et al., 2020a) are text-image alignment tasks, which focus on coarse-grained and fine-grained alignment, respectively.

However, the above methods rely on raster-scan serialization and may perform sub-optimally. Besides, with the conventional attention mechanism, the text, image and layout can not be fully interacted.

3 Approach

The conceptual overview of ERNIE-Layout is shown in Figure 2. Given a document image, incorporating the layout-knowledge of the document extracted from the Document Parser, ERNIE-Layout rearranges the segment (token) sequence in the order which is more consistent with human reading habits. We extract visual embeddings from Visual Encoder. We combine the textual embeddings and the layout embeddings into the textual feature through a linear projection, and similar operations are conducted for the visual feature. The textual and visual features are concatenated and fed into the Transformer layers, which utilize our new spatial-aware disentangled attention mechanism. For pre-training, ERNIE-Layout adopts 4 pre-training tasks, consisting of our newly proposed Reading Order Prediction, Replaced Region Prediction, and the traditional Masked Visual-Language Modeling, Text-Image Alignment.

In this section, we first introduce the Document-Parser module. Next, we describe how to get the input representation. Then, the multi-modal Transformer based on spatial-aware disentangled attention is described. Finally, we introduce the pre-training tasks used in ERNIE-Layout.

3.1 Document-Parser

The OCR is a commonly used module for VDU. Through OCR, we can obtain the textual words and their position coordinates in the document. The conventional methods arrange these words directly in the raster-scan order as the preprocessing step.

This method can't handle documents with complex layout properly, although it is easy to implement. As the example shown in figure 1, for information extraction from a given table, the expected value is a cell across multiple lines. Following the raster-scan order, the value to be extracted will contain lines of other cells, resulting in an incorrect prediction. This situation is more common in the cases with complex layout, such as multi-column paper, magazine, bill and report. Therefore, we use the Document-Parser, which can rearrange the textual words according to the layout-knowledge, and benefits the following multi-modal modeling.

The Document-Parser is a commercial layout analysis toolkit[1]. It can parse the document into different parts with their layouts according to the spatial distribution of words, pictures and tables, with a case in point is illustrated in Figure 2.

To evaluate the benefits of Document-Parser, we use PPL as the evaluation metric, which is widely used for evaluating the performance of language models. We calculate PPL by GPT-2 (Radford et al., 2019) to evaluate the quality of the process of token sequence. We find the token sequences serialized by Document-Parser obtain a lower PPL compared with those in the raster-scan order, and it tends to more significant for the document with complex layout. More implementation details and cases are shown in Appendix A.1.

3.2 Input Representation

The input features of ERNIE-Layout include textual feature and visual feature. The feature of each modality is the combination of its embeddings and the corresponding layout embeddings.

Text Embedding: The document tokens processed by Document-Parser module are used as the text sequence. To get the text embeddings, following BERT (Devlin et al., 2018), the special tokens $[CLS]$ and $[SEP]$ are concatenated at the beginning and end of the text sequence, respectively. Besides, a series of the $[PAD]$ tokens are appended after the last $[SEP]$ to ensure each token sequence length is the same length. In this way, the text embeddings $\boldsymbol{T}$ can be expressed as:

$$\boldsymbol{T} = E_{token}(\boldsymbol{T}^*) + E_{pos}(\boldsymbol{T}^*) + E_{type}(\boldsymbol{T}^*),$$

where $\boldsymbol{T}^*$ is the padded text sequence, E_{token} represents the text embedding layer, E_{pos} denotes the 1D position embedding layer, and E_{type} is the token type embedding layer. The length of text embeddings is L.

Visual Embedding: The document image is resized to 224×224. We use the Faster-RCNN (Ren et al., 2015) as the backbone and take the feature map of the second block. And then, we use an adaptive pooling layer to resize the feature map to $\mathbb{R}^{C \times H \times W}$, the typical values in our experiment are $C = 256, H = 7, W = 7$. We flatten the feature map into a sequence, and use a linear projection layer to map the visual sequence to the same dimension as the text embeddings. Similar to the method of processing text, image sequence is also fused with its 1D position and token type embeddings. Therefore, the visual embeddings $\boldsymbol{V}$ can be represented as:

$$\boldsymbol{V} = FC(\boldsymbol{V}^*) + E_{pos}(\boldsymbol{V}^*) + E_{type}(\boldsymbol{V}^*),$$

where $\boldsymbol{V}^*$ is the flattened visual sequence. And the length of visual embeddings is $H \times W$

Layout Embedding: For the textual sequence, following LayoutLM (Xu et al., 2020b), the token 2D position $(x_0, y_0, x_2, y_2, w, h)$ output by OCR are used as the layout information, where the (x_0, y_0) is the coordinates of the upper left corner, the (x_2, y_2) is the coordinates of the bottom right corner, and $w = x_2 - x_0$, $h = y_2 - y_0$, all the position values are normalized in the range [0, 1000]. The spatial information of special tokens $[CLS]$, $[SEP]$, $[PAD]$ are defined as $(0, 0, 0, 0, 0, 0)$. For visual sequence, similar spatial coordinates can also be obtained. We use separate embedding layers to get the layout vectors in the horizontal and vertical directions respectively, and the layout embeddings can be expressed as:

$$\boldsymbol{L} = E_x([\boldsymbol{T}^*; \boldsymbol{V}^*]) + E_y([\boldsymbol{T}^*; \boldsymbol{V}^*]),$$

where the E_x is the x-axis embedding layer, the E_y denotes the y-axis embedding layer. The length of layout embeddings is $L + HW$

To obtain the final input features $\boldsymbol{S}$ for ERNIE-Layout, the text embeddings and visual embeddings are fused with their corresponding layout embeddings, and are concatenated together, which can be represented as

$$\boldsymbol{S} = [\boldsymbol{W}; \boldsymbol{V}] + \boldsymbol{L}$$

3.3 Multi-Modal Transformer

We use an encoder-only Transformer to model the concatenated sequence $\boldsymbol{S}$ of the textual and visual features for a joint representation. To calculate the attention weights between tokens with respect to embeddings and their spatial information, we propose spatial-aware disentangled attention, which utilizing 1D and 2D relative position simultaneously. The 1D relative distance between token i and j is calculated by function δ_p as follows:

$$\delta_p(i,j) = \begin{cases} 0 & \text{for } \ i - j \leqslant -k \\ 2k - 1 & \text{for } \ i - j \geqslant k \\ i - j + k & \text{others}, \end{cases}$$

[1]https//anonymous.com/Document-Parser

where k is the maximum relative distance and the defined distance above can also be used for the 2D. $\boldsymbol{P}^r$, $\boldsymbol{X}^r$, $\boldsymbol{Y}^r \in \mathbb{R}^{2k\times d}$ represent relative position embedding layers, where d is the hidden size of Transformer. The projection matrices $\boldsymbol{W}^* \in \mathbb{R}^{d\times d}$ is used to generate the projected vectors $\boldsymbol{Q}^*$, $\boldsymbol{K}^*$ and $\boldsymbol{V}^*$ of content and relative position respectively, which can be obtained by the following expression:

$$\boldsymbol{Q}^c = \boldsymbol{S}'\boldsymbol{W}^{qc}, \boldsymbol{K}^c = \boldsymbol{S}'\boldsymbol{W}^{kc}, \boldsymbol{V}^c = \boldsymbol{S}'\boldsymbol{W}^{vc},$$

$$\boldsymbol{Q}^p = \boldsymbol{P}^r\boldsymbol{W}^{qp}, \boldsymbol{K}^p = \boldsymbol{P}^r\boldsymbol{W}^{kp},$$

$$\boldsymbol{Q}^x = \boldsymbol{X}^r\boldsymbol{W}^{qx}, \boldsymbol{K}^x = \boldsymbol{X}^r\boldsymbol{W}^{kx},$$

$$\boldsymbol{Q}^y = \boldsymbol{Y}^r\boldsymbol{W}^{qy}, \boldsymbol{K}^y = \boldsymbol{Y}^r\boldsymbol{W}^{ky},$$

where $\boldsymbol{S}'$ is the input vectors of the Transformer layer.

Besides the content attention matrix $\hat{A}^{cc}_{ij} = \boldsymbol{Q}^c_i {\boldsymbol{K}^c_j}^T$, we also calculate the attention bias between the content and relative position which can be expressed as:

$$\hat{A}^{cp}_{ij} = \boldsymbol{Q}^c_i {\boldsymbol{K}^p_{\delta_p(i,j)}}^T + \boldsymbol{K}^c_j {\boldsymbol{Q}^p_{\delta_p(j,i)}}^T,$$

$$\hat{A}^{cx}_{ij} = \boldsymbol{Q}^c_i {\boldsymbol{K}^x_{\delta_x(i,j)}}^T + \boldsymbol{K}^c_j {\boldsymbol{Q}^x_{\delta_x(j,i)}}^T,$$

$$\hat{A}^{cy}_{ij} = \boldsymbol{Q}^c_i {\boldsymbol{K}^y_{\delta_y(i,j)}}^T + \boldsymbol{K}^c_j {\boldsymbol{Q}^y_{\delta_y(j,i)}}^T$$

Finally, all these attention scores are summed up to get $\hat{\boldsymbol{A}}$. We apply a scaling factor of $1/3$ on $\hat{\boldsymbol{A}}$, which is important for stabilizing training. So, the output of spatial-aware disentangled attention module is:

$$\boldsymbol{H}_o = Softmax(\frac{\hat{\boldsymbol{A}}}{\sqrt{3d}})\boldsymbol{V}$$

Compared to previous methods, it avoids premature fusion of different types of relative position information.

3.4 Pre-training Tasks

Reading Order Prediction: The OCR results consist of several segments, which contain the tokens together with the corresponding layouts within them. However, there is no explicit boundary between segments in the sequence which is processed by Transformer. To enhance the token interactions within segments and correlation between segments, we propose Reading Order Prediction. We use vanilla self-attention to calculate token-level attention matrix, where the attention score represents the probability of the target token being the next token of the source token. The golden label of target token is the real next token. While the last token in segment points to itself, the other tokens point to the next token along the reading order. The loss of this task is:

$$\mathcal{L}_{ROP} = -\sum_{i\in L}\sum_{j\in L} A^{gt}_{ij} log(A^{pre}_{ij}),$$

where golden matrix $\boldsymbol{A}^{gt}$ contains the one-hot ground truth labels, and the prediction matrix A^{pre} contains the calculated probabilities.

Replaced Regions Prediction: Since the textual content is highly aligned with the image content in VDU task, the conventional image-text matching task modeling the alignment following the whole image-text level. The completely irrelevant image and text tend to be too simple for the model to classify. So, we propose Replaced Regions Prediction, which is a fine-grained multi-modal matching task. First of all, the original image will be defined into $H \times W$ patches, where the H, W are consistent with the corresponding values of the pooling layer after Visual Encoder. And we replace each patch with random region from another image with a probability of 10%. Then, the processed image will be encoded by the visual encoder and input into the Transformer. Finally, the $[CLS]$ vector output by Transformer will be used to predict which patches were replaced. So the loss of this task can be expressed as:

$$\mathcal{L}_{RPP} = -\sum_{i\in HW} [I^{gt}_i log(I^p_i) + (1-I^{gt}_i) log(1-I^p_i)],$$

where I^{gt} is the golden label of replaced patches, I^p indicates the normalized probability of predict logit.

Moreover, the conventional Masked Visual-Language Modeling and Text-Image Alignment pre-training tasks are also implemented in ERNIE-Layout, the final pre-training loss is represented as:

$$\mathcal{L} = \mathcal{L}_{ROP} + \mathcal{L}_{RRP} + \mathcal{L}_{MVLM} + \mathcal{L}_{TIA}$$

4 Experiments

4.1 Pre-training Details

For the pre-training dataset, similar to LayoutLM, we crawl the homologous data of the IIT-CDIP Test Collection (Lewis et al., 2006) from

Dataset	Key Number	Train	Dev	Test
FUNSD	4	149	0	50
CORD	30	800	100	100
SROIE	4	626	0	347
Kleister-NDA	4	254	83	203
RVL-CDIP	16	320K	40K	40K
DocVQA	-	39K	5K	5K

Table 1: Statistics of datasets for downstream tasks

Dataset	Epoch	Weight Decay	Batch
FUNSD	100	0	2
CORD	30	0.05	16
SROIE	100	0.05	16
Kleister-NDA	30	0.05	16
RVL-CDIP	20	0.05	16
DocVQA	6	0.05	16

Table 2: Hyper-parameters for downstream tasks

Tabacco website [2], which contains over 30 million scanned document pages. For a fair comparison with previous works, we randomly select 10 million pages as the pre-training dataset, and extract texts, layouts and word-level bounding boxes with Document-Parser.

For the Transformer architecture, we use 24 Transformer layers with 1024 hidden units and 16 heads. The maximum sequence length of text tokens and image block tokens are 512 and 49 respectively. The Transformer is initialized from RoBERTa (Liu et al., 2019) and Visual Encoder use the backbone of Faster-RCNN (Ren et al., 2015) as the initialized model. The rest parameters are randomly initialized.

We use Adam (Kingma and Ba, 2014) as the optimizer, with a learning rate of 1e-4 and a weight decay of 0.01. The learning rate is linearly warmed up over the first 10% steps then linearly decayed to 0. ERNIE-Layout is trained on 24 A100 GPUs for 20 epochs with a batch size of 576.

4.2 Downstream Tasks

We carry out experiments for Information Extraction tasks on FUNSD (Jaume et al., 2019), CORD (Park et al., 2019), SROIE (Biten et al., 2019), Kleister-NDA (Graliński et al., 2020), Document Question Answering task (DocVQA (Mathew et al., 2021)) and Document Classification task on RVL-CDIP (Harley et al., 2015). Table 1 shows the brief statistics of these fine-tuning datasets and more details about them are shown in Appendix A.2.

We solve Information Extraction tasks (FUNSD, CORD, SROIE, Kleister-NDA) in a sequence labeling manner and use a token-level classification layer to predict the BIO labels. For the Document Question Answering task (DocVQA), we use an extractive question-answering paradigm and build a token-level classifier after the ERNIE-Layout output representation to predict the start and end position of the answer. For the Document Classification task (RVL-CDIP), the representation of $[CLS]$ is processed by a fully-connected network to predict the document label.

For all the downstream tasks, we fine-tune ERNIE-Layout using Adam optimizer, with a learning rate of 2e-5, weight decay of 0.01. The learning rate is linearly warmed up and then linearly decayed. Other hyper-parameters are shown in Table 2. All the experiments are conducted on A100 GPUs.

4.3 Experimental Results

Table 3 shows the results for Information Extraction task on all the four datasets, which we use entity level F1 score to evaluate the abilities of the models. ERNIE-Layout achieves SOTA results on FUNSD, CORD, Kleister-NDA datasets. Especially in the FUNSD, ERNIE-Layout obtains a great improvement of 7.98% compared with the previous best results. ERNIE-Layout also achieves an improvement of 1.20%, 2.90% on CORD, Kleister-NDA respectively. The above results show that our model is superior to the existing multi-modal methods for Information Extraction task.

Table 4 shows the Average Normalized Levenshtein Similarity (ANLS) scores on the DocVQA dataset. Compared with the text-only baselines and previous best performing multi-modal models, our method achieves comparable result. While TILT, StructralLM don't clearly describe Fine-tuning set, we conduct thorough comparisons with LayoutLMv2. The results #2 and #3 show that, UniLMv2$_{\text{large}}$ is 7.57% higher than RoBERTa$_{\text{large}}$. Since UniLMv2 $_{\text{large}}$ doesn't expose model's code and parameters, we use RoBERTa$_{\text{large}}$ as the initialization parameter. The results of $\triangle$ANLS in #7b and #8b show that ERNIE-Layout$_{\text{large}}$($\triangle$ANLS:0.1534) is more significant than LayoutLMv2$_{\text{large}}$($\triangle$ANLS:0.0820). The improvement shows the effectiveness of our model. Finally, we achieve top-1 on the DocVQA

[2]https://www.industrydocuments.ucsf.edu/tobacco/

Method	FUNSD F1	CORD F1	SROIE F1	Kleister-NDA F1
$\text{BERT}_{\text{large}}$ (Liu et al., 2019)	0.6563	0.9025	0.9200	0.7910
$\text{RoBERTa}_{\text{large}}$ (Liu et al., 2019)	0.7072	-	0.9280	-
$\text{UniLMv2}_{\text{large}}$ (Bao et al., 2020)	0.7257	0.9205	0.9488	0.8180
$\text{LayoutLM}_{\text{large}}$ (Xu et al., 2020b)	0.7895	0.9493	0.9524	0.8340
$\text{TILT}_{\text{large}}$ (Powalski et al., 2021)	-	0.9633	0.9810	-
$\text{LayoutLMv2}_{\text{large}}$ (Xu et al., 2020a)	0.8420	0.9601	0.9781	0.8520
$\text{StructralLM}_{\text{large}}$ (Li et al., 2021a)	0.8514	-	-	-
$\text{ERNIE-Layout}_{\text{large}}$	**0.9312**	**0.9721**	0.9755	**0.8810**

Table 3: Results of ERNIE-Layout compared with previous methods for Information Extraction task

#	Method	Fine-tuning set	ANLS	△ANLS
1	$\text{BERT}_{\text{large}}$ (Liu et al., 2019)	train	0.6768	
2	$\text{RoBERTa}_{\text{large}}$ (Liu et al., 2019)	train	0.6952	
3	$\text{UniLMv2}_{\text{large}}{}^{\dagger}$ (Bao et al., 2020)	train	0.7709	
4	$\text{LayoutLM}_{\text{large}}$ (Xu et al., 2020b)	train	0.7808	
5	$\text{TILT}_{\text{large}}$ (Powalski et al., 2021)	-	0.8705	
6	$\text{StructralLM}_{\text{large}}$ (Li et al., 2021a)	-	0.8349	
7a	$\text{LayoutLMv2}_{\text{large}}{}^{\dagger}$ (Xu et al., 2020a)	train	0.8348	
7b	$\text{LayoutLMv2}_{\text{large}}$	train + dev	0.8529	0.0820
8a	$\text{ERNIE-Layout}_{\text{large}}$	train	0.8321	
8b	$\text{ERNIE-Layout}_{\text{large}}$	train+dev	0.8486	**0.1534**
9	$\text{ERNIE-Layout}_{\text{large}}$(leaderboard)	train+dev	**0.8841**	

Table 4: Results of ERNIE-Layout compared with previous methods for Document Question Answering task. "-" means Fine-tuning set not clearly described in origin paper. △ANLS means ANLS difference between text-only model and multi-modal model initialized from the corresponding text-only model, where ERNIE-Layout is based on RoBERTa and LayoutLMv2 is based on UniLMv2.

leaderboard by ensembling.

4.4 Ablation Study

Serialization Module	FUNSD F1	CORD F1
w. serialization in the raster-scan order	0.9128	0.9658
w. serialization by Document-Parser	0.9171	0.9678

Table 5: Ablation study on the FUNSD and CORD datasets of different serialization modules. Serialization in the raster-scan order means serialization by conventional OCR, and serialization by Document-Parser means rearranging the tokens with layout-knowledge.

We conduct ablation experiments to fully study the benefits of incorporating layout-knowledge, the proposed pre-training tasks and the spatial-aware disentangled attention mechanism. We use the same hyper-parameters settings for all the experiments and pre-train the models for 5 epochs. We use FUNSD and CORD datasets for the performance evaluation.

Effectiveness of incorporating layout-knowledge: We serialize the document into tokens following the raster-scan order and layout-knowledge enhanced order, respectively. This is the only difference for the pre-training. As the results shown in Table 5, serialization by Document-Parser is better than serialization in the raster-scan order with an improvement of 0.5% on FUNSD, which prove the effectiveness of incorporating layout-knowledge.

Effectiveness of the proposed pre-training tasks: We implement the baselines with the pre-training tasks MVLM and TIA from LayoutLMv2. Based on the baselines, we additionally adopt our newly proposed RRP and ROP. The experimental results are shown in Table 6. The RRP brings an improvement of 0.95% and 0.10% on FUNSD and CORD respectively, which shows the benefit of the fine-grained text-image alignment. Further

#	SADAM	SASAM	MVLM	TIA	RRP	ROP	FUNSD F1	CORD F1
1			√				0.8712	0.9513
2			√	√			0.8753	0.9555
3			√	√	√		0.8848	0.9565
4			√	√	√	√	0.8978	0.9603
5		√	√	√	√	√	0.9128	0.9658
6	√		√	√	√	√	0.9241	0.9673

Table 6: Ablation study on the FUNSD and CORD datasets. "SADAM" means the spatial-aware disentangled attention mechanism. "SASAM" means the spatial-aware self-attention mechanism. "MVLM", "TIA" are proposed pre-training tasks by LayoutLMv2. "RRP" and "ROP" are the two preposed pre-training tasks by our model.

Method	Accuracy
$\text{BERT}_{\text{large}}$ (Liu et al., 2019)	89.92%
$\text{RoBERTa}_{\text{large}}$ (Liu et al., 2019)	90.11%
$\text{UniLMv2}_{\text{large}}$ (Bao et al., 2020)	90.20%
$\text{LayoutLM}_{\text{large}}$ (Xu et al., 2020b)	94.43%
$\text{TILT}_{\text{large}}$ (Powalski et al., 2021)	95.52%
$\text{LayoutLMv2}_{\text{large}}$ (Xu et al., 2020a)	95.64%
$\text{StructralLM}_{\text{large}}$ (Li et al., 2021a)	96.08%
$\text{ERNIE-Layout}_{\text{large}}$	95.41%

Table 7: Results of ERNIE-Layout compared with previous methods for Document Classification task.

utilizing of ROP, brings a great improvement of 1.3% on FUNSD (#3 vs #4). We consider that ROP forces the model to build the joint representation containing more segment-level information.

Effectiveness of the spatial-aware disentangled attention mechanism: While the SADAM is an improved version of SASAM, we conduct experiments to study the benefit. From the results shown in Table 6, compared with SASAM, the model with SADAM achieves an improvement of 1.13% on FUNSD (#6 vs #5), which indicates that, our newly proposed attention mechanism helps to build better interaction between text-image feature and spatial feature.

4.5 Discussion

We get superior performance on Information Extraction and Question Answering tasks, which shows the effectiveness of our proposed method. For document classification, ERNIE-Layout also achieves comparable results and an improvement of 0.98% compared with LayoutLM, as shown in Table 7. But there is still a performance gap between ERNIE-Layout and the best model for this task. We consider the reasons are two folds. We use RoBERTa as our initialization model, which is less competitive compared with UniLMv2 used in LayoutLMv2 and T5 (Raffel et al., 2019) used in TILT. On the other hand, our pre-training tasks are designed for fine-grained document understanding and cross-modal alignment, which plays a less crucial role for Document Understanding.

5 Conclusion

In this work, we present ERNIE-Layout, the first layout-knowledge enhanced document pre-training approach to improve the performance of pre-training model in document understanding. ERNIE-Layout attempts to rearrange the parsed tokens from the document according to the layout-knowledge from Document Parser, and obtain a considerable improvement over the conventional raster-scan order. We propose the Reading Order Prediction task to force the model to build the joint representation containing more segment-level information. Furthermore, we propose a fine-grained text-image alignment task, Replace Region Prediction. We design a new attention mechanism to help to build better interaction between text-image feature and spatial feature. The extensive experiments demonstrate the effectiveness of our proposed method. While ERNIE-Layout hasn't achieved the best result for Document Classification, for future work, we will attempt to enhance the document level modeling during the pre-training process.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer:

End-to-end transformer for document understanding. *arXiv preprint arXiv:2106.11539*.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE.

Mengli Cheng, Minghui Qiu, Xing Shi, Jun Huang, and Wei Lin. 2020. One-shot text field labeling using attention and belief propagation for structure information extraction. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 340–348.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. Lambert: Layout-aware language modeling for information extraction. In *International Conference on Document Analysis and Recognition*, pages 532–547. Springer.

Filip Graliński, Tomasz Stanisławek, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2020. Kleister: A novel task for information extraction involving long documents with complex layout. *arXiv preprint arXiv:2003.02356*.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

G. Jaume, H. K. Ekenel, and J. P. Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. *IEEE*.

Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. *arXiv preprint arXiv:1809.08799*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021a. Structurallm: Structural pre-training for form understanding. *arXiv preprint arXiv:2105.11210*.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021b. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660.

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*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209.

Rasmus Berg Palm, Florian Laws, and Ole Winther. 2019. Attend, copy, parse end-to-end information extraction from documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 329–336. IEEE.

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 *Workshop on Document Intelligence at NeurIPS 2019*.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. *arXiv preprint arXiv:2102.09550*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Clément Sage, Alex Aussem, Véronique Eglin, Haytham Elghazel, and Jérémy Espinas. 2020. End-to-end extraction of structured information from business documents with pointer-generator networks. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 43–52.

Ritesh Sarkhel and Arnab Nandi. 2019. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. In *28th International Joint Conference on Artificial Intelligence (IJCAI), 2019*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2738–2745.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

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

Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. 2017. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

A Appendix

A.1 The Effects of Document-Parser

The Document-Parser assembles multiple modules such as document-specific OCR, Layout Parser, and Table Parser. The Layout Parser and Table Parser module play a crucial role for the incorporation of layout-knowledge in ERNIE-Layout.

An important preprocessing step for the document understanding is serializing the extracted document tokens. The popular method for this serialization is performed directly on the output results of OCR in raster-scan order and is suboptimal though simple to implement. With the Layout Parser and Table Parser of the Document Parser toolkit, the order of the tokens will be further rearranged according to the layout-knowledge. During the parsing processing, the tables and figures will be detected as spatial layouts, and the free texts will be processed by paragraph analysis which combines heuristics and detection models to get the paragraph layout information and the upper-lower boundary relationship.

Parallel Session A		
Fire Dynamics 1	**Risk 1**	**Evacuation**
Room: ACM15 1.001	Room: ACM15 1.008	Room: FKJ 12 0.06
Session Chair: *Tuula Hakkarainen*	*Session Chair:* ***Frank Markert***	*Session Chair:*

Figure 3: The example used to show the difference between serialization method. The serialization by the raster-scan order is "... Session Chair: Session Chair: Session Chair: Tuula Hakkarainen ...". And the serialization by Document-Parser is "... Session Chair: Tuula Hakkarainen Session Chair: Frank Markert ...", which is more consistent with human reading habits.

An example is shown in Figure 3, which is extracted from the third image in table 8 is used to show the sequence serialized by the raster-scan order and Document-Parser, respectively.

To validate the effectiveness of our method, we use an open-sourced language model GPT-2 (Wolf et al., 2020), to calculate the PPL of the serialized token sequence by the raster-scan order and Document-Parser respectively. Since documents with complex layouts only account for a small proportion of the total documents, in a test of 10,000 documents, the average PPL only drops about 1 point, but on documents with complex layouts, as shown in 8, Document-Parser shows great advantages.

A.2 Details of Fine-tuning Datasets

FUNSD (Jaume et al., 2019) is a dataset for form understanding on noisy scanned documents that aims at extracting values from forms. FUNSD comprises 199 real, fully annotated, scanned forms. The training set contains 149 samples, and the test set contains 50 samples. We use the official OCR annotations. Following previous methods, we adopt the entity-level F1 score as the evaluation metric. Similar to StructralLM (Li et al., 2021a), we use the cell-level layout information when performing the fine-tuning.

CORD (Park et al., 2019) is a consolidated dataset for receipt parsing as the first step towards post-OCR parsing tasks. CORD consists of thousands of Indonesian receipts, which contain images and box/text annotations for OCR, and multi-level semantic labels for parsing. The training set, validation set, and test set contain 800, 100, and 100 receipts respectively. We use the official OCR annotations and the entity-level F1 score as the evaluation metric.

SROIE (Huang et al., 2019) is a scanned receipts OCR and key information extraction dataset, which covers important aspects related to the automated analysis of scanned receipts. The training set and test set contain 626 and 347 samples respectively. This task requires the model to extract values from each receipt of four predefined keys: company, date, address, and total. We use the official OCR annotations and the entity-level F1 score as the evaluation metric.

Kleister-NDA (Graliński et al., 2020) is provided for key information extraction task, which involves a mix of scanned and born-digital long formal documents. The training set, valid set, and test set contain 254, 83, 203 samples respectively. Due to that the test set is not publicly available, we report the entity-level F1 score on the validation set, which is computed by the official evaluation tools[3].

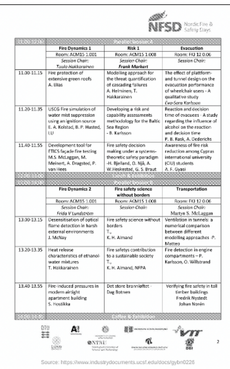
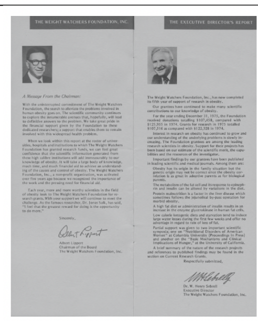
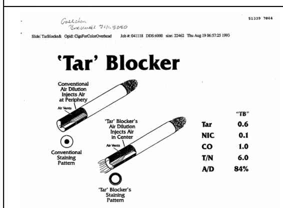
Document Page	RSO	DP
	100.39	67.98
	98.99	42.02
	146.66	76.87
	70.12	25.61
	219.47	170.54

Table 8: The PPL results of serialized token sequence according to different methods. RSO denotes the raster-scan order and DP indicates the Document-Parser

[3]https://gitlab.com/filipg/geval

The task aims to extract values of four predefined keys: date, jurisdiction, party, and term.

RVL-CDIP (Harley et al., 2015) is a document classification dataset consisting of grayscale document images. The training set, validation set, and test set contain 320000, 40000, and 40000 document images respectively. The document images are categorized into 16 classes, with 25000 images per class. We use Microsoft OCR tools to extract text and layout information from document images, and the evaluation metric is classification accuracy.

DocVQA (Mathew et al., 2021) is a dataset for Visual Question Answering (VQA) on document images. The dataset consists of 50000 questions defined on 12767 document images. The document images are split into the training set, validation set, and test set with the ratio of 8:1:1. We use the Microsoft OCR tools to extract the texts and layouts from document images. The task aims to predict the start and end position of the answer span. ANLS (average normalized Levenshtein similarity) (Biten et al., 2019) is used as the evaluation metric.