LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding

Jiapeng Wang[1] **Lianwen Jin***[1,3,4] **Kai Ding***[2,3]
[1]South China University of Technology, Guangzhou, China
[2]IntSig Information Co., Ltd, Shanghai, China
[3]INTSIG-SCUT Joint Laboratory of Document Recognition and Understanding, China
[4]Peng Cheng Laboratory, Shenzhen, China
[1]eejpwang@mail.scut.edu.cn, eelwjin@scut.edu.cn
[2]danny_ding@intsig.net

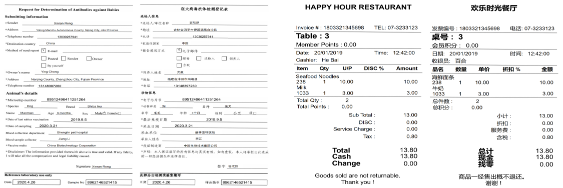
Abstract

Structured document understanding has attracted considerable attention and made significant progress recently, owing to its crucial role in intelligent document processing. However, most existing related models can only deal with the document data of specific language(s) (typically English) included in the pre-training collection, which is extremely limited. To address this issue, we propose a simple yet effective **L**anguage-**i**ndependent **L**ayout **T**ransformer (**LiLT**) for structured document understanding. LiLT can be pre-trained on the structured documents of a single language and then directly fine-tuned on other languages with the corresponding off-the-shelf monolingual/multilingual pre-trained textual models. Experimental results on eight languages have shown that LiLT can achieve competitive or even superior performance on diverse widely-used downstream benchmarks, which enables language-independent benefit from the pre-training of document layout structure. Code and model are publicly available at https://github.com/jpWang/LiLT.

1 Introduction

Structured document understanding (SDU) aims at reading and analyzing the textual and structured information contained in scanned/digital-born documents. With the acceleration of the digitization process, it has been regarded as a crucial part of intelligent document processing and required by many real-world applications in various industries such as finance, medical treatment and insurance.

Recently, inspired by the rapid development of pre-trained language models of plain texts (Devlin et al., 2019; Liu et al., 2019b; Bao et al., 2020; Chi et al., 2021), many researches on structured document pre-training (Xu et al., 2020, 2021a,b; Li et al., 2021a,b,c; Appalaraju et al., 2021) have also pushed the limit of a variety of SDU tasks. However, almost all of them only focus on pre-training and fine-tuning on the documents in a single language, typically English. This is extremely limited for other languages, especially in the case of lacking pre-training structured document data.



Request for Determination of Antibodies against Rabies

Submitting information

*Sender Xinran Rong
*Address Yitong Manchu Autonomous County, Siping City, Jilin Province
*Telephone number 13030257941
*Destination country China
*Method of send report ☑ E-mail
□ Posted □ Sender □ Owner
□ By yourself
*Owner's name Ying Chong
*Address Nanjing County, Zhangzhou City, Fujian Province
*Telephone number 13148397260

Animal's details

*Microchip number 89512496411251264
*Species Dog Breed Shiba Inu
Name Maomao Age 3 months Sex Male☑ Female□
*Date of last rabies vaccination 2019.9.5
*Date of sampling 2020.3.21
Blood collection department Shenglin pet hospital
Blood sample collector Jiang Li
*Vaccine make China Biotechnology Corporation
*Disclaimer: The information provided therewith above is true and valid. If any falsity, I will take all the compensation and legal liability caused.

Signature Xinran Rong

Reference laboratory use only

Date 2020.4.26 Sample No 8962146521415

狂犬病毒抗体检测登记表

送检人信息

*送检人/单位名称 容欣然
*地址 吉林省四平市伊通满族自治县
*电话 13030257941
*拟前往国家 中国
*报告递送方式 ☑电子邮件
□邮寄 □送检人 □饲养人
□自取
*饲养人姓名 充融
*地址 福建省漳州市南靖县
*电话 13148397260

动物信息

*电子芯片号 89512496411251264
*动物种类 狗 品种 柴犬
名字 毛毛 年龄 3个月 性别 公☑ 母□
*最后免疫日期 2019.9.5
*采血日期 2020.3.21
采血单位 盛林宠物医院
采血人姓名 李江
*疫苗制造商 中国生物技术集团公司
*声明：本人保证填写的所有信息均真实有效，如有虚假，本人将承担由此造成的一切经济损失和法律责任。

签字 容欣然

此部分由检测实验室填写

日期 2020.4.26 样品编号 8962146521415

HAPPY HOUR RESTAURANT

Invoice # : 1803321345698 TEL: 07-3233123
Table : 3
Member Points : 0.00
Date: 20/01/2019 Time: 12:42:00
Cashier: He Bai

Item	Qty	U/P	DISC %	Amount
Seafood Noodles				
238	1	10.00		10.00
Milk				
1033	1	3.00		3.00

Total Qty : 2
Total Points : 0.00

Sub Total :	13.00
DISC :	0.00
Service Charge :	0.00
Tax :	0.80
Total	13.80
Cash	13.80
Change	0.00

Goods sold are not returnable.
Thank you !

欢乐时光餐厅

发票编号：1803321345698 电话: 07-3233123
桌号： 3
会员积分： 0.00
日期: 20/01/2019 时间: 12:42:00
收银员: 百合

品名	数量	单价	折扣 %	金额
海鲜面条				
238	1	10.00		10.00
牛奶				
1033	1	3.00		3.00

总件数： 2
总积分： 0.00

小计：	13.00
折扣：	0.00
服务费：	0.00
含税：	0.80
总计	13.80
现金	13.80
找零	0.00

商品一经售出概不退还。
谢谢！

(a) A form. (b) A receipt.

Figure 1: The substitution of language does not appear obviously unnatural when the layout structure remains unchanged, as shown in a (a) form/(b) receipt. The detailed content has been re-synthesized to avoid the sensitive information leak. Best viewed in zoomed-in.

In this regard, we consider how to make the SDU tasks enjoy language-independent benefit from the pre-training of document layout structure. Here, we give an observation as shown in Figure 1. When the layout structure remains unchanged, the substitution of language does not make obvious unnaturalness. It fully motivates us to decouple and reuse the layout invariance among different languages.

Based on this inspiration, in this paper, we propose a simple yet effective **L**anguage-**i**ndependent **L**ayout **T**ransformer (**LiLT**) for structured document understanding. In our framework, the text and layout information are first decoupled and joint-optimized during pre-training, and then re-coupled for fine-tuning. To ensure that the two modalities have sufficient language-independent interaction, we further propose a novel bi-directional attention complementation mechanism (BiACM) to enhance the cross-modality cooperation. Moreover, we present the key point location (KPL) and cross-modal alignment identification (CAI) tasks, which are combined with the widely-used masked visual-

[*] Corresponding author.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7747 - 7757
May 22-27, 2022 ©2022 Association for Computational Linguistics

language modeling (MVLM) to serve as our pre-training objectives. During fine-tuning, the layout flow (LiLT) can be separated and combined with the off-the-shelf pre-trained textual models (such as RoBERTa (Liu et al., 2019b), XLM-R (Conneau et al., 2020), InfoXLM (Chi et al., 2021), etc) to deal with the downstream tasks. In this way, our method decouples and learns the layout knowledge from the monolingual structured documents before generalizing it to the multilingual ones.

To the best of our knowledge, the only pre-existing multilingual SDU model is LayoutXLM (Xu et al., 2021b). It scraps multilingual PDF documents of 53 languages from a web crawler and introduces extra pre-processing steps to clean the collected data, filter the low-quality documents, and classify them into different languages. After this, it utilizes a heuristic distribution to sample 22 million multilingual documents, which are further combined with the 8 million sampled English ones from the IIT-CDIP (Lewis et al., 2006) dataset (11 million English documents), resulting 30 million for pre-training with the LayoutLMv2 (Xu et al., 2021a) framework. However, this process is time-consuming and laborious. On the contrary, LiLT can be pre-trained with only IIT-CDIP and then adapted to other languages. In this respect, LiLT is the first language-independent method for structured document understanding.

Experimental results on eight languages have shown that LiLT can achieve competitive or even superior performance on diverse widely-used downstream benchmarks, which substantially benefits numerous real-world SDU applications. Our main contributions can be summarized as follows:

- We introduce a simple yet effective language-independent layout Transformer called LiLT for monolingual/multilingual structured document understanding.
- We propose BiACM to provide language-independent cross-modality interaction, along with an effective asynchronous optimization strategy for textual and non-textual flows in pre-training. Moreover, we present two new pre-training objectives, namely KPL and CAI.
- LiLT achieves competitive or even superior performance on various widely-used downstream benchmarks of different languages under different settings, which fully demonstrates its effectiveness.

2 LiLT

Figure 2 shows the overall illustration of our method. Given an input document image, we first use off-the-shelf OCR engines to get text bounding boxes and contents. Then, the text and layout information are separately embedded and fed into the corresponding Transformer-based architecture to obtain enhanced features. Bi-directional attention complementation mechanism (BiACM) is introduced to accomplish the cross-modality interaction of text and layout clues. Finally, the encoded text and layout features are concatenated and additional heads are added upon them, for the self-supervised pre-training or the downstream fine-tuning.

2.1 Model Architecture

The whole framework can be regarded as a parallel dual-stream Transformer. The layout flow shares a similar structure as text flow, except for the reduced hidden size and intermediate size to achieve computational efficiency.

2.1.1 Text Embedding

Following the common practice (Devlin et al., 2019; Xu et al., 2020), in the text flow, all text strings in the OCR results are first tokenized and concatenated as a sequence S_t by sorting the corresponding text bounding boxes from the top-left to bottom-right. Intuitively, the special tokens `[CLS]` and `[SEP]` are also added at the beginning and end of the sequence respectively. After this, S_t will be truncated or padded with extra `[PAD]` tokens until its length equals the maximum sequence length N. Finally, we sum the token embedding E_{token} of S_t and the 1D positional embedding P_{1D} to obtain the text embedding $E_T \in \mathcal{R}^{N\times d_T}$ as:

$$E_T = \mathrm{LN}(E_{token} + P_{\mathrm{1D}}), \tag{1}$$

where d_T is the number of text feature dimension and LN is the layer normalization (Ba et al., 2016).

2.1.2 Layout Embedding

As for the layout flow, we construct a 2D position sequence S_l with the same length as the token sequence S_t using the corresponding text bounding boxes. To be specific, we normalize and discretize all box coordinates to integers in the range $[0, 1000]$, and use four embedding layers to generate x-axis, y-axis, height, and width features separately. Given the normalized bounding boxes $B = (x_{min}, x_{max}, y_{min}, y_{max}, width, height)$, the 2D

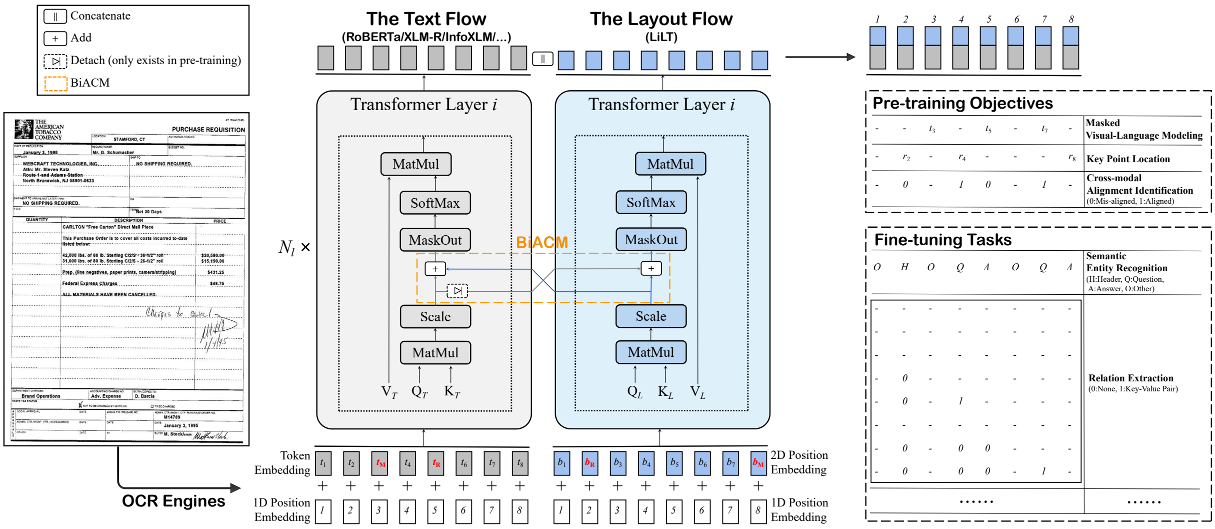


Figure 2: The overall illustration of our framework. Text and layout information are separately embedded and fed into the corresponding flow. BiACM is proposed to accomplish the cross-modality interaction. At the model output, text and layout features are concatenated for the self-supervised pre-training or the downstream fine-tuning. N_l is the number of Transformer layers. The red $*_{\mathrm{M}}$/$*_{\mathrm{R}}$ indicates the randomly masked/replaced item for pre-training. t, b and r represent *token*, *box* and *region*, respectively. Best viewed in zoomed-in.

positional embedding $P_{\text{2D}} \in \mathcal{R}^{N \times d_L}$ (where d_L is the number of layout feature dimension) is constructed as follows:

$$P_{\text{2D}} = Linear(\text{CAT}(E_{x_{min}}, E_{x_{max}}, E_{y_{min}}, E_{y_{max}}, E_{width}, E_{height})). \quad (2)$$

Here, the Es are embedded vectors. $Linear$ is a linear projection layer and CAT is the channel-wise concatenation operation. The special tokens `[CLS]`, `[SEP]` and `[PAD]` are also attached with (0,0,0,0,0,0), (1000,1000,1000,1000,0,0) and (0,0,0,0,0,0) respectively. It is worth mentioning that, for each token, we directly utilize the bounding box of the text string it belongs to, because the fine-grained token-level information is not always included in the results of some OCR engines.

Since Transformer layers are permutation-invariant, here we introduce the 1D positional embedding again. The resulting layout embedding $E_L \in \mathcal{R}^{N \times d_L}$ can be formulated as:

$$E_L = \text{LN}(P_{\text{2D}} + P_{\text{1D}}). \quad (3)$$

2.1.3 BiACM

The text embedding E_T and layout embedding E_L are fed into their respective sub-models to generate high-level enhanced features. However, it will considerably ignore the cross-modal interaction process if we simply combine the text and layout features at the encoder output only. The network also needs to comprehensively analyse them at earlier stages. In view of this, we propose a new bi-directional attention complementation mechanism (BiACM) to strengthen the cross-modality interaction across the entire encoding pipeline. Experiments in Section 3.2 will further verify its effectiveness.

The vanilla self-attention mechanism in Transformer layers captures the correlation between query x_i and key x_j by projecting the two vectors and calculating the attention score as:

$$\alpha_{ij} = \frac{(x_i W^Q)(x_j W^K)^\top}{\sqrt{d^h}}. \quad (4)$$

Here, the description is for a single head in a single self-attention layer with hidden size of d^h and projection metrics W^Q, W^K for simplicity. Given α_{ij}^T and α_{ij}^L of the text and layout flows located in the same head of the same layer, BiACM shares them as common knowledge, which is formulated as:

$$\widetilde{\alpha_{ij}^T} = \alpha_{ij}^L + \alpha_{ij}^T, \quad (5)$$

$$\widetilde{\alpha_{ij}^L} = \begin{cases} \alpha_{ij}^L + \text{DETACH}(\alpha_{ij}^T) & if\ \text{Pre-train}, \\ \alpha_{ij}^L + \alpha_{ij}^T & if\ \text{Fine-tune}. \end{cases} \quad (6)$$

In order to maintain the ability of LiLT to cooperate with different off-the-shelf text models in fine-tuning as much as possible, we heuristically adopt the detached α_{ij}^T for $\widetilde{\alpha_{ij}^L}$, so that the textual stream will not be affected by the gradient of non-textual one during pre-training, and its overall consistency can be preserved. Finally, the modified attention scores are used to weight the projected value vectors for subsequent modules in both flows.

2.2 Pre-training Tasks

We conduct three self-supervised pre-training tasks to guide the model to autonomously learn joint representations with cross-modal cooperation. The details are introduced below.

2.2.1 Masked Visual-Language Modeling

This task is originally derived from (Devlin et al., 2019). MVLM randomly masks some of the input tokens and the model is asked to recover them over the whole vocabulary using the output encoded features, driven by a cross-entropy loss. Meanwhile, the non-textual information remains unchanged. MVLM improves model learning on the language side with cross-modality information. The given layout embedding can also help the model better capture both inter- and intra-sentence relationships. We mask 15% text tokens, among which 80% are replaced by the special token `[MASK]`, 10% are replaced by random tokens sampled from the whole vocabulary, and 10% remain the same.

2.2.2 Key Point Location

We propose this task to make the model better understand layout information in the structured documents. KPL equally divides the entire layout into several regions (we set 7×7=49 regions by default) and randomly masks some of the input bounding boxes. The model is required to predict which regions the key points (top-left corner, bottom-right corner, and center point) of each box belong to using separate heads. To deal with it, the model is required to fully understand the text content and know where to put a specific word/sentence when the surrounding ones are given. We mask 15% boxes, among which 80% are replaced by (0,0,0,0,0,0), 10% are replaced by random boxes sampled from the same batch, and 10% remain the same. Cross-entropy loss is adopted.

Since there may exist detection errors in the output of OCR engines, we let the model predict the discretized regions (as mentioned above) instead of the exact location. This strategy can moderately relax the punishment criterion while improving the model performance.

2.2.3 Cross-modal Alignment Identification

We collect those encoded features of token-box pairs that are masked and further replaced (misaligned) or kept unchanged (aligned) by MVLM and KPL, and build an additional head upon them to identify whether each pair is aligned. To achieve this, the model is required to learn the cross-modal perception capacity. CAI is a binary classification task, and a cross-entropy loss is applied for it.

2.3 Optimization Strategy

Utilizing a unified learning rate for all model parameters to perform the end-to-end training process is the most common optimization strategy. While in our case, it will cause the layout flow to continuously update in the direction of coupling with the evolving text flow in the pre-training stage, which is harmful to the ability of LiLT to cooperate with different off-the-shelf textual models during fine-tuning. Based on this consideration, we explore multiple ratios to greatly slow down the pre-training optimization of the text stream. We also find that an appropriate reduction ratio is better than parameter freezing.

Note that, we adopt a unified learning rate for end-to-end optimization during fine-tuning. The DETACH operation of BiACM is also canceled at this time, as shown in Equation 6.

3 Experiments

3.1 Pre-training Setting

We pre-train LiLT on the IIT-CDIP Test Collection 1.0 (Lewis et al., 2006), which is a large-scale scanned document image dataset and contains more than 6 million documents with more than 11 million scanned document images. We use TextIn API[1] to obtain the text bounding boxes and strings for this dataset.

In this paper, we initialize the text flow from the existing pre-trained English RoBERTa$_{\text{BASE}}$ (Liu et al., 2019b) for our document pre-training, and combine LiLT$_{\text{BASE}}$ with the pre-trained InfoXLM$_{\text{BASE}}$ (Chi et al., 2021)/a new pre-trained RoBERTa$_{\text{BASE}}$ for multilingual/monolingual fine-tuning. They have an equal number of self-attention layers, attention heads and maximum sequence length, which ensures that BiACM can work normally. In this BASE setting, LiLT has a 12-layer encoder with 192 hidden size, 768 feed-forward filter size and 12 attention heads, resulting

[1] https://www.textin.com

#	Inter-modal Operation	Average F1
1	CAT	0.6751
2	CAT+Co-Attention (Lu et al., 2019)	0.6276
3	CAT+BiACM	**0.7963**
4	CAT+BiACM−DETACH in pre-training	0.7682
5	CAT+BiACM+DETACH in fine-tuning	0.7822
6	The text flow alone (InfoXLM$_{\text{BASE}}$, as shown in Table 6)	0.7207

(a) BiACM. CAT is short for concatenation.

#	MVLM	KPL	CAI	Average F1
1	✓			0.7616
2	✓	✓		0.7748
3	✓		✓	0.7809
4	✓	✓	✓	**0.7963**

(b) Pre-training tasks.

#	Slow-down Ratio	Average F1
1	1 (No Slow-down)	0.7840
2	500	0.7901
3	800	0.7947
4	1000	**0.7963**
5	1200	0.7935
6	$+\infty$ (Parameter Freezing)	0.7893

(c) Slow-down ratios.

Table 1: Ablation study of LiLT$_{\text{BASE}}$ combined with InfoXLM$_{\text{BASE}}$ (Chi et al., 2021) on the FUNSD and XFUND datasets (8 languages in total). The average F1 accuracy of language-specific semantic entity recognition (SER) task is given. (a) BiACM. (b) Pre-training tasks. (c) Slow-down ratios of the pre-training optimization for the text flow.

in the number of parameters as 6.1M. The maximum sequence length N is set as 512.

LiLT$_{\text{BASE}}$ is pre-trained using Adam optimizer (Kingma and Ba, 2015; Loshchilov and Hutter, 2018), with the learning rate 2×10^{-5}, weight decay 1×10^{-2}, and (β_1, β_2) = (0.9, 0.999). The learning rate is linearly warmed up over the first 10% steps and then linearly decayed. We set the batch size as 96 and train LiLT$_{\text{BASE}}$ for 5 epochs on the IIT-CDIP dataset using 4 NVIDIA A40 48GB GPUs.

3.2 Ablation Study

Considering that the complete pre-training takes a relatively long time, we pre-train LiLT$_{\text{BASE}}$ with 2M documents randomly sampled from IIT-CDIP for 5 epochs to conduct ablation experiments, as shown in Table 1.

Model	Precision	Recall	F1
BERT$_{\text{BASE}}$[1]	0.5469	0.6710	0.6026
RoBERTa$_{\text{BASE}}$[2]	0.6349	0.6975	0.6648
UniLMv2$_{\text{BASE}}$[3]	0.6349	0.6975	0.6648
LayoutLM$_{\text{BASE}}$[4]	0.7597	0.8155	0.7866
BROS$_{\text{BASE}}$[5]	0.8056	0.8188	0.8121
SelfDoc[6]	-	-	0.8336
LayoutLMv2$_{\text{BASE}}$[7]	0.8029	0.8539	0.8276
StrucText$_{\text{BASE}}$[8]	0.8568	0.8097	0.8309
DocFormer$_{\text{BASE}}$[9]	0.8076	0.8609	0.8334
⋆LayoutXLM$_{\text{BASE}}$[10]	0.7913	0.8158	0.8034
LiLT[EN-R[2]]$_{\text{BASE}}$	**0.8721**	**0.8965**	**0.8841**
⋆**LiLT**[InfoXLM[11]]$_{\text{BASE}}$	0.8467	0.8709	0.8586

Table 2: Comparison on the semantic entity recognition (SER) task of FUNSD (Jaume et al., 2019) dataset. **Bold** indicates the SOTA and underline indicates the second best. "EN-R" is short for English RoBERTa. ⋆The multilingual model. **[]** denotes the off-the-shelf textual model used as the text flow of LiLT. [1](Devlin et al., 2019);[2](Liu et al., 2019b);[3](Bao et al., 2020);[4](Xu et al., 2020);[5](Hong et al., 2020);[6](Li et al., 2021b);[7](Xu et al., 2021a);[8](Li et al., 2021c);[9](Appalaraju et al., 2021);[10](Xu et al., 2021b);[11](Chi et al., 2021).

We first evaluate the effect of introducing BiACM. In setting (a)#1, the text and layout features are concatenated at the model output without any further interaction. Compared with (a)#6, we find that such a plain design results in a much worse performance than using the text flow alone. From (a)#1 to (a)#3, the significant improvement demonstrates that it is the novel BiACM that makes the transfer from "monolingual" to "multilingual" successful. Beside this, we have also tried to replace BiACM with the co-attention mechanism (Lu et al., 2019) which is widely adopted in dual-stream Transformer architecture. It can be seen as a "deeper" cross-modal interaction, since the keys and values from each modality are passed as input to the other modality's dot-product attention calculation. However, severe drops are observed as shown in (a)#2 vs (a)#1#3. We attribute it to the damage of such a "deeper" interaction to the overall consistency of the text flow in the pre-training optimization. In contrast, BiACM can maintain LiLT's cross-model cooperation ability on the basis of providing cross-modal information. Moreover, the necessity of DETACH in pre-training is proved in (a)#4 vs (a)#3. Compared (a)#3 to (a)#5, we can also infer that removing DETACH in fine-tuning leads to a better performance.

Then, we compare the proposed KPL and CAI tasks. As shown in Table 1(b), both tasks improve the model performance substantially, and the proposed CAI benefits the model more than KPL. Using both tasks together is more effective than using either one alone.

Model	Precision	Recall	F1
BERT$_{\text{BASE}}$	0.8833	0.9107	0.8968
UniLMv2$_{\text{BASE}}$	0.8987	0.9198	0.9092
LayoutLM$_{\text{BASE}}$	0.9437	0.9508	0.9472
BROS$_{\text{BASE}}$	0.9558	0.9514	0.9536
LAMBERT$_{\text{BASE}}$[1]	-	-	0.9441
TILT$_{\text{BASE}}$[2]	-	-	0.9511
LayoutLMv2$_{\text{BASE}}$	0.9453	0.9539	0.9495
DocFormer$_{\text{BASE}}$	**0.9652**	0.9614	**0.9633**
*LayoutXLM$_{\text{BASE}}$	0.9456	0.9506	0.9481
LiLT[EN-R]$_{\text{BASE}}$	0.9598	**0.9616**	0.9607
***LiLT**[InfoXLM]$_{\text{BASE}}$	0.9574	0.9581	0.9577

Table 3: Comparison on the semantic entity recognition (SER) task of CORD (Park et al., 2019) dataset. [1](Garncarek et al., 2021);[2](Powalski et al., 2021).

Model	Precision	Recall	F1
BiLSTM+CRF[1]	-	-	0.8910
GraphIE[2]	-	-	0.9026
GCN-based[3]	-	-	0.9255
TRIE[4]	-	-	0.9321
VIES[5]	-	-	0.9523
MatchVIE[6]	-	-	0.9687
TCPN[7]	-	-	0.9759
RoBERTa$_{\text{BASE}}$[8]	0.9405	0.9640	0.9521
StrucTexT$_{\text{BASE}}$	-	-	0.9795
*LayoutXLM$_{\text{BASE}}$	0.9699	0.9820	0.9759
LiLT[ZH-R[8]]$_{\text{BASE}}$	**0.9762**	**0.9833**	**0.9797**
***LiLT**[InfoXLM]$_{\text{BASE}}$	0.9699	0.9820	0.9759

Table 4: Comparison on the semantic entity recognition (SER) task of EPHOIE (Wang et al., 2021a) dataset. "ZH-R" is short for Chinese RoBERTa. [1](Lample et al., 2016);[2](Qian et al., 2019);[3](Liu et al., 2019a);[4](Zhang et al., 2020);[5](Wang et al., 2021a);[6](Tang et al., 2021);[7](Wang et al., 2021b);[8](Cui et al., 2020).

Model	Accuracy
VGG-16[1]	90.97%
Stacked CNN Single[2]	91.11%
Stacked CNN Ensemble[2]	92.21%
InceptionResNetV2[3]	92.63%
LadderNet[4]	92.77%
Multimodal Single[5]	93.03%
Multimodal Ensemble[5]	93.07%
BERT$_{\text{BASE}}$	89.81%
UniLMv2$_{\text{BASE}}$	90.06%
LayoutLM$_{\text{BASE}}$ (w/ image)	94.42%
BROS$_{\text{BASE}}$	95.58%
SelfDoc	93.81%
TILT$_{\text{BASE}}$	93.50%
LayoutLMv2$_{\text{BASE}}$	95.25%
DocFormer$_{\text{BASE}}$	**96.17%**
*LayoutXLM$_{\text{BASE}}$	95.21%
LiLT[EN-R]$_{\text{BASE}}$	95.68%
***LiLT**[InfoXLM]$_{\text{BASE}}$	95.62%

Table 5: Comparison on the document classification (DC) task of RVL-CDIP (Harley et al., 2015) dataset. [1](Afzal et al., 2017);[2](Das et al., 2018);[3](Szegedy et al., 2017);[4](Sarkhel and Nandi, 2019);[5](Dauphinee et al., 2019).

Finally, we explore the most suitable slow-down ratio for the pre-training optimization of the text flow. A ratio equal to 1 in (c)#1 means there is no slow-down and a unified learning rate is adopted. It can be found that the F1 scores keep rising with the growth of slow-down ratios and begin to fall when the ratio is greater than 1000. Consequently, we set the slow-down ratio as 1000 by default.

3.3 Comparisons with the SOTAs

To demonstrate the performance of LiLT, we conduct experiments on several widely-used monolingual datasets and the multilingual XFUND benchmark (Xu et al., 2021b). In addition to the experiments involving typical language-specific fine-tuning, we also follow the two settings designed in (Xu et al., 2021b) to demonstrate the ability to transfer knowledge among different languages, which are zero-shot transfer learning and multitask fine-tuning, for fair comparisons. Specifically, (1) language-specific fine-tuning refers to the typical fine-tuning paradigm of fine-tuning on language X and testing on language X. (2) Zero-shot transfer learning means the models are fine-tuned on English data only and then evaluated on each target language. (3) Multitask fine-tuning requires the model to fine-tune on data in all languages.

3.3.1 Language-specific Fine-tuning

We first evaluate LiLT on four widely-used monolingual datasets - FUNSD (Jaume et al., 2019), CORD (Park et al., 2019), EPHOIE (Wang et al., 2021a) and RVL-CDIP (Lewis et al., 2006), and the results are shown in Table 2, 3, 4 and 5. We have found that (1) LiLT is flexible since it can work with monolingual or multilingual plain text models to deal with downstream tasks. (2) Although LiLT is designed for the transfer from "monolingual" to "multilingual", it can surprisingly cooperate with monolingual textual models to achieve competitive or even superior performance (especially on the FUNSD dataset with only a few training samples available), compared with existing language-specific SDU models such as LayoutLMv2 and

Task	Model	Pre-training Docs		FUNSD	XFUND							Avg.
		Language	Size	EN	ZH	JA	ES	FR	IT	DE	PT	
SER	XLM-RoBERTa$_{\text{BASE}}$	-	-	0.6670	0.8774	0.7761	0.6105	0.6743	0.6687	0.6814	0.6818	0.7047
	InfoXLM$_{\text{BASE}}$	-	-	0.6852	0.8868	0.7865	0.6230	0.7015	0.6751	0.7063	0.7008	0.7207
	LayoutXLM$_{\text{BASE}}$	Multilingual	30M	0.7940	0.8924	0.7921	0.7550	0.7902	0.8082	0.8222	0.7903	0.8056
	LiLT[InfoXLM]$_{\text{BASE}}$	**English only**	**11M**	**0.8415**	**0.8938**	**0.7964**	**0.7911**	**0.7953**	**0.8376**	**0.8231**	**0.8220**	**0.8251**
RE	XLM-RoBERTa$_{\text{BASE}}$	-	-	0.2659	0.5105	0.5800	0.5295	0.4965	0.5305	0.5041	0.3982	0.4769
	InfoXLM$_{\text{BASE}}$	-	-	0.2920	0.5214	0.6000	0.5516	0.4913	0.5281	0.5262	0.4170	0.4910
	LayoutXLM$_{\text{BASE}}$	Multilingual	30M	0.5483	0.7073	0.6963	0.6896	0.6353	0.6415	0.6551	0.5718	0.6432
	LiLT[InfoXLM]$_{\text{BASE}}$	**English only**	**11M**	**0.6276**	**0.7297**	**0.7037**	**0.7195**	**0.6965**	**0.7043**	**0.6558**	**0.5874**	**0.6781**

Table 6: Language-specific fine-tuning F1 accuracy on FUNSD and XFUND (fine-tuning on X, testing on X). “SER” denotes the semantic entity recognition and “RE” denotes the relation extraction. [] indicates the off-the-shelf textual model used as the text flow of LiLT.

DocFormer. (3) On these datasets which are widely adopted for monolingual evaluation, LiLT generally performs better than LayoutXLM. This fully demonstrates the effectiveness of our pre-training framework and indicates that the layout and text information can be successfully decoupled in pre-training and re-coupled in fine-tuning.

Then we evaluate LiLT on language-specific fine-tuning tasks of FUNSD and the multilingual XFUND (Xu et al., 2021b), and the results are shown in Table 6. Compared with the plain text models (XLM-R/InfoXLM) or the LayoutXLM model pre-trained with 30M multilingual structured documents, LiLT achieves the highest F1 scores on both the SER and RE tasks of each language while using 11M monolingual data. This significant improvement shows LiLT’s capability to transfer language-independent knowledge from pre-training to downstream tasks.

3.3.2 Zero-shot Transfer Learning

The results of cross-lingual zero-shot transfer are presented in Table 7. It can be observed that the LiLT model transfers the most knowledge from English to other languages, and significantly outperforms its competitors. This fully verifies that LiLT can capture the common layout invariance among different languages. Moreover, LiLT has never seen non-English documents before evaluation under this setting, while the LayoutXLM model has been pre-trained with them. This is to say, LiLT faces a stricter cross-lingual zero-shot transfer scenario but achieves better performance.

3.3.3 Multi-task Fine-tuning

Table 8 shows the results of multitask learning. In this setting, the pre-trained LiLT model is simultaneously fine-tuned with all eight languages and evaluated for each specific language. We observe that this setting further improves the model performance compared to the language-specific fine-tuning, which confirms that SDU can benefit from commonalities in the layout of multilingual structured documents. In addition, LiLT once again outperforms its counterparts by a large margin.

4 Related Work

During the past decade, deep learning methods became the mainstream for document understanding tasks (Yang et al., 2017; Augusto Borges Oliveira et al., 2017; Siegel et al., 2018). Grid-based methods (Katti et al., 2018; Denk and Reisswig, 2019; Lin et al., 2021) were proposed for 2D document representation where text pixels were encoded using character or word embeddings and classified into specific field types, using a convolutional neural network. GNN-based approaches (Liu et al., 2019a; Yu et al., 2021; Tang et al., 2021) adopted multi-modal features of text segments as nodes to model the document graph, and used graph neural networks to propagate information between neighboring nodes to attain a richer representation.

In recent years, self-supervised pre-training has achieved great success. Inspired by the development of the pre-trained language models in various NLP tasks, recent studies on structured document pre-training (Xu et al., 2020, 2021a,b; Li et al., 2021a,b,c; Appalaraju et al., 2021) have pushed the limits. LayoutLM (Xu et al., 2020) modified the BERT (Devlin et al., 2019) architecture by adding 2D spatial coordinate embeddings. In comparison, our LiLT can be regarded as a more powerful and flexible solution for structured document understanding. LayoutLMv2 (Xu et al., 2021a) improved over LayoutLM by treating the visual fea-

Task	Model	Pre-training Docs		FUNSD	XFUND							Avg.
		Language	Size	EN	ZH	JA	ES	FR	IT	DE	PT	
SER	XLM-RoBERTa$_{\text{BASE}}$	-	-	0.6670	0.4144	0.3023	0.3055	0.3710	0.2767	0.3286	0.3936	0.3824
	InfoXLM$_{\text{BASE}}$	-	-	0.6852	0.4408	0.3603	0.3102	0.4021	0.2880	0.3587	0.4502	0.4119
	LayoutXLM$_{\text{BASE}}$	Multilingual	30M	0.7940	0.6019	0.4715	0.4565	0.5757	0.4846	0.5252	0.5390	0.5561
	LiLT[InfoXLM]$_{\text{BASE}}$♠	**English only**	**11M**	**0.8415**	**0.6152**	**0.5184**	**0.5101**	**0.5923**	**0.5371**	**0.6013**	**0.6325**	**0.6061**
RE	XLM-RoBERTa$_{\text{BASE}}$	-	-	0.2659	0.1601	0.2611	0.2440	0.2240	0.2374	0.2288	0.1996	0.2276
	InfoXLM$_{\text{BASE}}$	-	-	0.2920	0.2405	0.2851	0.2481	0.2454	0.2193	0.2027	0.2049	0.2423
	LayoutXLM$_{\text{BASE}}$	Multilingual	30M	0.5483	0.4494	0.4408	0.4708	0.4416	0.4090	0.3820	0.3685	0.4388
	LiLT[InfoXLM]$_{\text{BASE}}$♠	**English only**	**11M**	**0.6276**	**0.4764**	**0.5081**	**0.4968**	**0.5209**	**0.4697**	**0.4169**	**0.4272**	**0.4930**

Table 7: Cross-lingual zero-shot transfer F1 accuracy on FUNSD and XFUND (fine-tuning on FUNSD, testing on X). ♠ indicates that LiLT faces a stricter zero-shot transfer scenario compared with LayoutXLM, since it has never seen non-English documents before evaluation, even during pre-training.

Task	Model	Pre-training Docs		FUNSD	XFUND							Avg.
		Language	Size	EN	ZH	JA	ES	FR	IT	DE	PT	
SER	XLM-RoBERTa$_{\text{BASE}}$	-	-	0.6633	0.8830	0.7786	0.6223	0.7035	0.6814	0.7146	0.6726	0.7149
	InfoXLM$_{\text{BASE}}$	-	-	0.6538	0.8741	0.7855	0.5979	0.7057	0.6826	0.7055	0.6796	0.7106
	LayoutXLM$_{\text{BASE}}$	Multilingual	30M	0.7924	0.8973	0.7964	0.7798	0.8173	0.8210	0.8322	0.8241	0.8201
	LiLT[InfoXLM]$_{\text{BASE}}$	**English only**	**11M**	**0.8574**	**0.9047**	**0.8088**	**0.8340**	**0.8577**	**0.8792**	**0.8769**	**0.8493**	**0.8585**
RE	XLM-RoBERTa$_{\text{BASE}}$	-	-	0.3638	0.6797	0.6829	0.6828	0.6727	0.6937	0.6887	0.6082	0.6341
	InfoXLM$_{\text{BASE}}$	-	-	0.3699	0.6493	0.6473	0.6828	0.6831	0.6690	0.6384	0.5763	0.6145
	LayoutXLM$_{\text{BASE}}$	Multilingual	30M	0.6671	0.8241	0.8142	0.8104	0.8221	0.8310	0.7854	0.7044	0.7823
	LiLT[InfoXLM]$_{\text{BASE}}$	**English only**	**11M**	**0.7407**	**0.8471**	**0.8345**	**0.8335**	**0.8466**	**0.8458**	**0.7878**	**0.7643**	**0.8125**

Table 8: Multitask fine-tuning F1 accuracy on FUNSD and XFUND (fine-tuning on 8 languages all, testing on X).

tures as separate tokens. Furthermore, additional pre-training tasks were explored to improve the utilization of unlabeled document data. SelfDoc (Li et al., 2021b) established the contextualization over a block of content, while StructuralLM (Li et al., 2021a) proposed cell-level 2D position embeddings and the corresponding pre-training objective. Recently, StrucTexT (Li et al., 2021c) introduced a unified solution to efficiently extract semantic features from different levels and modalities to handle the entity labeling and entity linking tasks. DocFormer (Appalaraju et al., 2021) designed a novel multi-modal self-attention layer capable of fusing textual, vision and spatial features.

Nevertheless, the aforementioned SDU approaches mainly focus on a single language - typically English, which is extremely limited with respect to multilingual application scenarios. To the best of our knowledge, LayoutXLM (Xu et al., 2021b) was the only pre-existing multilingual SDU model, which adopted the multilingual textual model InfoXLM (Chi et al., 2021) as the initialization, and adapted the LayoutLMv2 (Xu et al., 2021a) framework to multilingual structured document pre-training. However, it required a heavy process of multilingual data collection, cleaning and pre-training. On the contrary, our LiLT can deal with the multilingual structured documents by pre-training on the monolingual IIT-CDIP Test Collection 1.0 (Lewis et al., 2006) only.

5 Conclusion

In this paper, we present LiLT, a language-independent layout Transformer that can learn the layout knowledge from monolingual structured documents and then generalize it to deal with multilingual ones. Our framework successfully first decouples the text and layout information in pre-training and then re-couples them for fine-tuning. Experimental results on eight languages under three settings (language-specific, cross-lingual zero-shot transfer, and multi-task fine-tuning) have fully illustrated its effectiveness, which substantially bridges the language gap in real-world structured document understanding applications. The public availability of LiLT is also expected to promote the development of document intelligence.

For future research, we will continue to follow the pattern of transferring from "monolingual" to "multilingual" and further unlock the power of LiLT. In addition, we will also explore the generalized rather than language-specific visual information contained in multilingual structured documents.

6 Acknowledgement

This research is supported in part by NSFC (Grant No.: 61936003) and GD-NSF (No. 2017A030312006).

References

Muhammad Zeshan Afzal, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. 2017. Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification. In *ICDAR*, volume 1, pages 883–888.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. DocFormer: End-to-end Transformer for document understanding. In *ICCV*.

Dario Augusto Borges Oliveira et al. 2017. Fast CNN-based document layout analysis. In *ICCV Workshop*, pages 1173–1180.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

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 *ICML*, pages 642–652.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *NAACL-HLT*, pages 3576–3588.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*, pages 8440–8451.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of EMNLP*, pages 657–668.

Arindam Das, Saikat Roy, Ujjwal Bhattacharya, and Swapan K Parui. 2018. Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In *ICPR*, pages 3180–3185.

Tyler Dauphinee, Nikunj Patel, and Mohammad Rashidi. 2019. Modular multimodal architecture for document classification. *arXiv preprint arXiv:1912.04376*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, and Filip Graliński. 2021. LAMBERT: Layout-aware (language) modeling using BERT for information extraction. In *ICDAR*.

Adam W Harley et al. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *ICDAR*, pages 991–995.

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2020. BROS: A pre-trained language model for understanding texts in document.

Guillaume Jaume et al. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *ICDAR*, volume 2, pages 1–6.

Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2D documents. In *EMNLP*, pages 4459–4469.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL-HLT*, pages 260–270.

David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *ACM SIGIR*, 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. In *ACL*.

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 *CVPR*, pages 5652–5660.

Yulin Li, Yuxi Qian, Yuchen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021c. StrucTexT: Structured text understanding with multi-modal Transformers. In *ACM-MM*.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125.

Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. 2021. ViBERTgrid: A jointly trained multi-modal 2D document representation for key information extraction from documents. In *ICDAR*.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019a. Graph convolution for multimodal information extraction from visually rich documents. In *NAACL-HLT*, pages 32–39.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *ICLR*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32:13–23.

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*.

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. In *ICDAR*.

Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 2019. GraphIE: A graph-based framework for information extraction. In *NAACL-HLT*, pages 751–761.

Ritesh Sarkhel and Arnab Nandi. 2019. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. In *IJCAI*, pages 3360–3366.

Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting scientific figures with distantly supervised neural networks. In *JCDL*, pages 223–232.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284.

Guozhi Tang, Lele Xie, Lianwen Jin, Jiapeng Wang, Jingdong Chen, Zhen Xu, Qianying Wang, Yaqiang Wu, and Hui Li. 2021. MatchVIE: Exploiting match relevancy between entities for visual information extraction. In *IJCAI*, pages 1039–1045.

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021a. Towards robust visual information extraction in real world: New dataset and novel solution. In *AAAI*, volume 35, pages 2738–2745.

Jiapeng Wang, Tianwei Wang, Guozhi Tang, Lianwen Jin, Weihong Ma, Kai Ding, and Yichao Huang. 2021b. Tag, copy or predict: A unified weakly-supervised learning framework for visual information extraction using sequences. In *IJCAI*, pages 1082–1090.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2021a. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *ACL*.

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 *ACM-SIGKDD*, pages 1192–1200.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021b. LayoutXLM: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*.

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 *CVPR*, pages 5315–5324.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2021. PICK: Processing key information extraction from documents using improved graph learning-convolutional networks. In *ICPR*, pages 4363–4370.

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-MM*, pages 1413–1422.

Appendix

A Dataset Details

FUNSD FUNSD (Jaume et al., 2019) is an English dataset for form understanding in noisy scanned documents. It contains 199 real, fully annotated, scanned forms where 9,707 semantic entities are annotated above 31,485 words. The 199 samples are split into 149 for training and 50 for testing. We directly use the official OCR annotations. The semantic entity recognition (SER) task is assigning to each word a semantic entity label from a set of four predefined categories: question, answer, header, or other. The entity-level F1 score is used as the evaluation metric (Table 2).

CORD CORD (Park et al., 2019) is an English receipt dataset for key information extraction. Its publicly available subset includes 800 receipts for the training set, 100 for the validation set, and 100 for the test set. A photo and a list of OCR annotations are equipped for each receipt. The dataset defines 30 fields under 4 categories and the task aims to label each word to the right field. The evaluation metric is the entity-level F1 score, as shown in Table 3. We use the official OCR annotations.

EPHOIE EPHOIE (Wang et al., 2021a) is collected from actual Chinese examination papers with the diversity of text types and layout distribution. The 1,494 samples are divided into a training set with 1,183 images and a testing set with 311 images, respectively. It defines ten entity categories, and we provide the entity-level F1 score for RoBERTa, LayoutXLM and LiLT in Table 4. The official OCR annotations are adopted.

RVL-CDIP RVL-CDIP (Harley et al., 2015) consists of 400,000 gray-scale images of English documents, with 8:1:1 for the training set, validation set, and test set. A multi-class single-label classification task is defined on RVL-CDIP. The images are categorized into 16 classes, with 25,000 images per class. The evaluation metric is the overall classification accuracy as shown in Table 5. Text and layout information are extracted by TextIn API.

XFUND XFUND (Xu et al., 2021b) is a multilingual form understanding dataset that contains 1,393 fully annotated forms with seven languages including Chinese (ZH), Japanese (JA), Spanish (ES), French (FR), Italian (IT), German (DE), and Portuguese (PT). Each language includes 199 forms, where the training set includes 149 forms, and the test set includes 50 forms. We focus on the semantic entity recognition (SER) and relation extraction (RE) tasks defined in the original paper (Xu et al., 2021b). Relation extraction aims to predict the relation between any two given semantic entities, and we mainly focus on the key-value relation extraction. We use the official OCR results, and the same F1 accuracy evaluation metric as in LayoutXLM (Xu et al., 2021b) for Table 6, 7 and 8.

B Fine-tuning Details

Fine-tuning for Semantic Entity Recognition We conduct the semantic entity recognition task on FUNSD, CORD, EPHOIE and XFUND. We build a token-level classification layer above the output representations to predict the BIO tags for each entity field.

Fine-tuning for Document Classification This task depends on high-level visual information, thereby we leverage the image features explicitly in the fine-tuning stage, following LayoutLMv2 (Xu et al., 2021a). We pool the visual feature of the ResNeXt101-FPN (Xie et al., 2017; Lin et al., 2017) backbone into a global feature, concatenate it with the [CLS] output feature, and feed them into the final classification layer.

Fine-tuning for Relation Extraction We build the additional head for relation extraction on the FUNSD and XFUND datasets following (Xu et al., 2021b) for fair comparison. We first incrementally construct the set of relation candidates by producing all possible pairs of given semantic entities. For every pair, the representation of the head/tail entity is the concatenation of the first token vector in each entity and the entity type embedding obtained with a specific type embedding layer. After respectively projected by two FFN layers, the representations of head and tail are concatenated and then fed into a bi-affine classifier.