XFUND: A Benchmark Dataset for Multilingual Visually Rich Form Understanding

Yiheng Xu[1]*, **Tengchao Lv**[1], **Lei Cui**[1], **Guoxin Wang**[2], **Yijuan Lu**[2], **Dinei Florencio**[2], **Cha Zhang**[2], **Furu Wei**[1]
[1]Microsoft Research Asia [2]Microsoft Azure AI
{t-yihengxu,tengchaolv,lecu}@microsoft.com
{guow,yijlu,dinei,chazhang,fuwei}@microsoft.com

Abstract

Multimodal pre-training with text, layout, and image has achieved SOTA performance for visually rich document understanding tasks recently, which demonstrates the great potential for joint learning across different modalities. However, the existed research work has focused only on the English domain while neglecting the importance of multilingual generalization. In this paper, we introduce a human-annotated multilingual form understanding benchmark dataset named **XFUND**, which includes form understanding samples in 7 languages (Chinese, Japanese, Spanish, French, Italian, German, Portuguese). Meanwhile, we present LayoutXLM, a multimodal pre-trained model for multilingual document understanding, which aims to bridge the language barriers for visually rich document understanding. Experimental results show that the LayoutXLM model has significantly outperformed the existing SOTA cross-lingual pre-trained models on the XFUND dataset. The XFUND dataset and pre-trained LayoutXLM models have been publicly available at https://aka.ms/layoutxlm.

1 Introduction

Recently, multimodal pre-training for visually rich document understanding (VRDU) has achieved new SOTA performance on several public benchmarks (Xu et al., 2021, 2020), including form understanding (Jaume et al., 2019), receipt understanding (Park et al., 2019), complex layout understanding (Stanisławek et al., 2021), document image classification (Harley et al., 2015) and document VQA task (Mathew et al., 2021), due to the advantage that text, layout and image information is jointly learned end-to-end in a single framework. However, since most evaluation benchmarks focus on English VRDs, it is hard to explore the performance of a document understanding system on VRDs in other languages. Simply translating these documents automatically with machine translation services might help, but it is often not satisfactory due to the poor translation quality on document images (Afli and Way, 2016). Therefore, it is vital to explore the multilingual generalization ability of multimodal pre-training for VRDU tasks.

Multilingual pre-trained models such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-RoBERTa (Conneau et al., 2020), mBART (Liu et al., 2020), and the recent InfoXLM (Chi et al., 2021) and mT5 (Xue et al., 2021) have pushed many SOTA results on cross-lingual natural language understanding tasks by pre-training the Transformer models on different languages. These models have successfully bridged the language barriers in a number of cross-lingual transfer benchmarks such as XNLI (Conneau et al., 2018) and XTREME (Hu et al., 2020). Although a large amount of multilingual text data has been used in these cross-lingual pre-trained models, text-only multilingual models cannot be easily used in the VRDU tasks because they are usually fragile in analyzing the documents due to the format/layout diversity of documents in different countries, and even different regions in the same country. Hence, to accurately understand these visually rich documents in different languages, it is crucial to pre-train the multilingual models in a multimodal framework. Meanwhile, it is vital to provide a human-labeled benchmark to further facilitate multilingual document understanding.

To this end, we introduce a human-annotated multilingual form understanding benchmark dataset named **XFUND**, which contains 7 languages, including Chinese, Japanese, Spanish, French, Italian, German, Portuguese. In addition to the fully annotated data, we propose two subtasks

*Contribution during internship at Microsoft Research Asia. Correspondence to Lei Cui<lecu@microsoft.com> and Furu Wei<fuwei@microsoft.com>

Findings of the Association for Computational Linguistics: ACL 2022, pages 3214 - 3224

with three different settings. The two subtasks are semantic entity recognition and relation extraction. And we introduce three different settings to explore the multilingual and complex layout generalization ability: (1) Language-specific fine-tuning follows the typical paradigm of fine-tuning and testing on the same language. (2) Zero-transfer learning means that the model is trained on English data only and then evaluated on each target language. (3) Multitask fine-tuning requires the model to be trained on data from all languages and then evaluated on each target language. These different settings evaluate not only the multilingual representation for each languages but also the cross-lingual generalization across tasks.

Moreover, we also present a multimodal pre-trained model for multilingual VRDU tasks, aka LayoutXLM, which is a multilingual extension of the recent LayoutLMv2 model (Xu et al., 2021). To evaluate the multilingual generalization ability of this framework, we use the pre-training objectives of LayoutLMv2, including Masked Visual-Language Model (MVLM), Image-Text Matching (ITM), and Image-Text Alignment (ITA). In addition, we pre-train the model with the IIT-CDIP dataset (Lewis et al., 2006) as well as a great number of publicly available digital-born multilingual PDF files from the internet, which helps the LayoutXLM model to learn from real-world documents. In this way, the model obtains textual and visual signals from a variety of document templates/layouts/formats in different languages, thereby taking advantage of the local invariance property from both textual, visual and linguistic perspectives. Experiment results show that the pre-trained LayoutXLM outperforms several SOTA cross-lingual pre-trained models(Conneau et al., 2020; Chi et al., 2021) on the XFUND benchmark dataset, which also demonstrates the potential of the multimodal pre-training strategy for multilingual document understanding.

The contributions of this paper are summarized as follows:

- We introduce XFUND, a multilingual form understanding benchmark dataset that includes human-labeled forms with key-value pairs in 7 languages (Chinese, Japanese, Spanish, French, Italian, German, Portuguese).
- We propose LayoutXLM, a multimodal pre-trained model for multilingual document understanding, which is trained with large-scale real-world scanned/digital-born documents.
- LayoutXLM has outperformed other SOTA multilingual baseline models on the XFUND dataset, which demonstrates the great potential for the multimodal pre-training for the multilingual VRDU task. The pre-trained LayoutXLM model and the XFUND dataset have been publicly available.

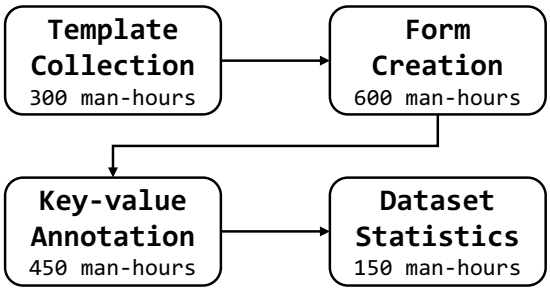


Figure 1: The illustration of corpus construction.

2 XFUND

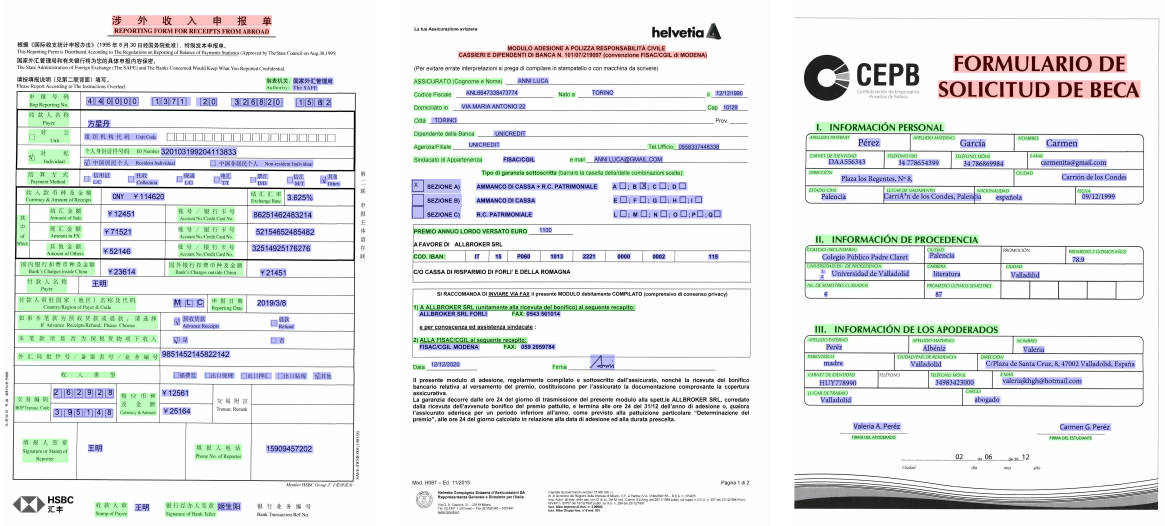
As illustrated in Figure 1, we develop our XFUND dataset in four steps including §2.1 Template Collection, §2.2 Form Creation, §2.3 Key-value Annotation, and §2.4 Data Finalization and Statistics, spending around 1,500 hours of human labor in total. Further details of ethic consideration are presented in §A Ethical Consideration.

2.1 Template Collection

Forms are usually used to collect information in different business scenarios. To avoid the privacy and sensitive information issue with real-world documents, we collect the documents publicly available on the internet and remove the content within the documents while only keeping the templates to fill in synthetic information manually. We collect form templates in 7 languages from the internet.

2.2 Form Creation

With the collected form templates, the human annotators manually fill synthetic information into these templates following corresponding requirements. Each template is allowed to be used only once, which means each form is different from the others. Besides, since the FUNSD (Jaume et al., 2019) documents contain both digitally filled-out forms and handwritten forms, we also ask annotators to fill in the forms by typing or handwriting. The completed



(a) Chinese (b) Italian (c) Spanish

Figure 2: Three sampled forms from the XFUND benchmark dataset (Chinese and Italian), where red denotes the headers, green denotes the keys and blue denotes the values.

forms are finally scanned into document images for further OCR processing and key-value labeling.

2.3 Key-value Annotation

Key-value pairs are also annotated by human annotators. Equipped with the synthetic forms, we use Microsoft Read API[1] to generate OCR tokens with bounding boxes. With an in-house GUI annotation tool, annotators are shown the original document images and the bounding boxes visualization of all OCR tokens. The annotators are asked to group the discrete tokens into entities and assign pre-defined labels to the entities. Also, if two entities are related, they are linked together as a key-value pair.

2.4 Data Finalization and Statistics

We design testing scripts to filter and check the annotated files and ask specific annotators for ethic checking. Cases with detected issues will be sent to the data annotation pipeline again for new valid labels.

Finally, the XFUND benchmark includes 7 languages with 1,393 fully annotated forms, where sampled documents are shown in Figure 2. Each language includes 199 forms, where the training set includes 149 forms, and the test set includes 50 forms. Detailed information is shown in Table 1.

[1] https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/overview-ocr

Lang	Split	Header	Question	Answer	Other	Total
ZH	training	229	3,692	4,641	1,666	10,228
	testing	58	1,253	1,732	586	3,629
JA	training	150	2,379	3,836	2,640	9,005
	testing	58	723	1,280	1,322	3,383
ES	training	253	3,013	4,254	3,929	11,449
	testing	90	909	1,218	1,196	3,413
FR	training	183	2,497	3,427	2,709	8,816
	testing	66	1,023	1,281	1,131	3,501
IT	training	166	3,762	4,932	3,355	12,215
	testing	65	1,230	1,599	1,135	4,029
DE	training	155	2,609	3,992	1,876	8,632
	testing	59	858	1,322	650	2,889
PT	training	185	3,510	5,428	2,531	11,654
	testing	59	1,288	1,940	882	4,169

Table 1: Statistics of the XFUND dataset. Each number in the table indicates the number of entities in each category.

2.5 Task Definition

Key-value extraction is one of the most critical tasks in form understanding. Inspired by FUNSD (Jaume et al., 2019), we define this task with two sub-tasks, which are semantic entity recognition and relation extraction.

Semantic Entity Recognition Given a visually rich document $\mathcal{D}$, we acquire discrete token set $t = \{t_0, t_1, ..., t_n\}$, where each token $t_i = (w, (x_0, y_0, x_1, y_1))$ consists of a word w and its bounding box coordinates (x_0, y_0, x_1, y_1).

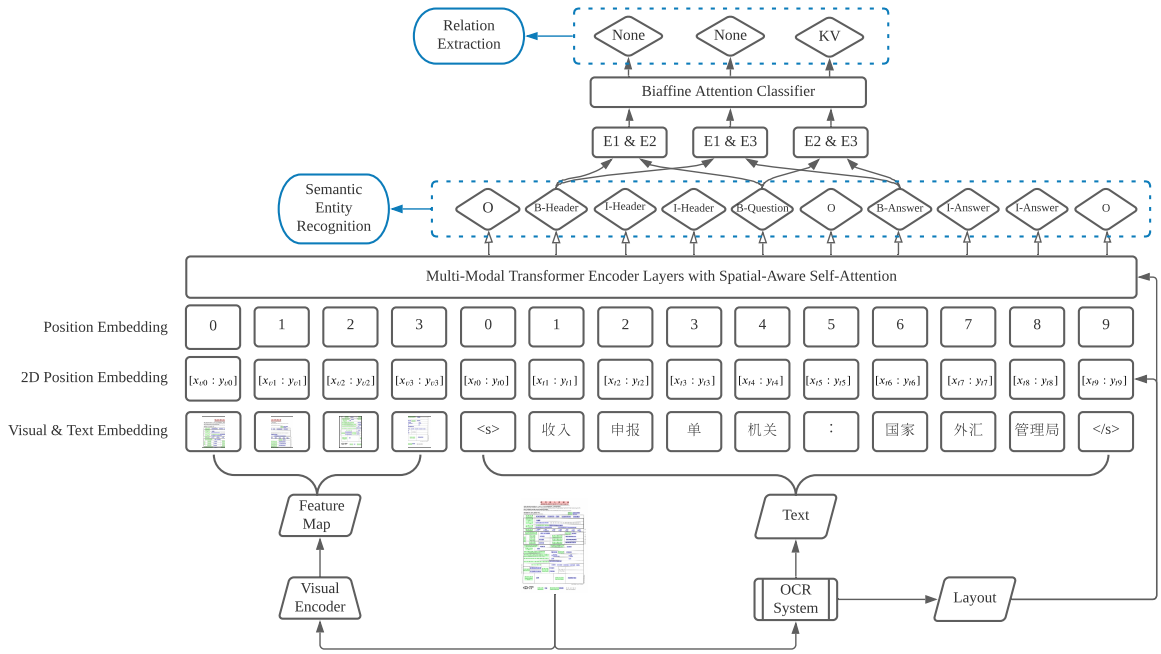


Figure 3: Architecture of the LayoutXLM Model, where the semantic entity recognition and relation extraction tasks are also demonstrated.

$\mathcal{C} = \{c_0, c_1, .., c_m\}$ is the semantic entity labels where the tokens are classified into. Semantic entity recognition is the task of extracting semantic entities and classifying them into given entity types. In other words, we intend to find a function $F_{SER} : (\mathcal{D}, \mathcal{C}) \rightarrow \mathcal{E}$, where $\mathcal{E}$ is the predicted semantic entity set:

$$\mathcal{E} = \{(\{t_0^0, ..., t_0^{n_0}\}, c_0), ..., (\{t_k^0, ..., t_k^{n_k}\}, c_k)\}$$

Relation Extraction Equipped with the document $\mathcal{D}$ and the semantic entity label set $\mathcal{C}$, relation extraction aims to predict the relation between any two predicted semantic entities. Defining $\mathcal{R} = \{r_0, r_1, .., r_m\}$ as the semantic relation labels, we intend to find a function $F_{RE} : (\mathcal{D}, \mathcal{C}, \mathcal{R}, \mathcal{E}) \rightarrow \mathcal{L}$, where $\mathcal{L}$ is the predicted semantic relation set:

$$\mathcal{L} = \{(head_0, tail_0, r_0), ..., (head_k, tail_k, r_k)\}$$

where $head_i$ and $tail_i$ are two semantic entities. In this work, we mainly focus on the key-value relation extraction.

3 LayoutXLM

In this section, we present a powerful baseline model LayoutXLM and introduce its model architecture, pre-training objectives, and pre-training dataset. We follow the LayoutLMv2 (Xu et al., 2021) architecture and transfer the model to large-scale multilingual document datasets.

3.1 Model Architecture

Similar to the LayoutLMv2 framework, we built the LayoutXLM model with a multimodal Transformer architecture. The framework is shown in Figure 3. The model accepts information from three different modalities, including text, layout, and image, which are encoded respectively with text embedding, layout embedding, and visual embedding layers. The text and image embeddings are concatenated, then plus the layout embedding to get the input embedding. The input embeddings are encoded by a multimodal Transformer with the spatial-ware self-attention mechanism. Finally, the output contextual representation can be utilized for the following task-specific layers. For brevity, we refer to (Xu et al., 2021) for further details on architecture.

3.2 Pre-training

The pre-training objectives of LayoutLMv2 have shown effectiveness in modeling visually rich documents. Therefore, we naturally adapt this pre-training framework to multilingual document pre-training. Following the idea of cross-modal alignment, our pre-training framework for docu-

ment understanding contains three pre-training objectives, which are Multilingual Masked Visual-Language Modeling (text-layout alignment), Text-Image Alignment (fine-grained text-image alignment), and Text-Image Matching (coarse-grained text-image alignment).

Multilingual Masked Visual-Language Modeling The Masked Visual-Language Modeling (MVLM) is originally proposed in the vanilla LayoutLM and also used in LayoutLMv2, aiming to model the rich text in visually rich documents. In this pre-training objective, the model is required to predict the masked text token based on its remaining text context and whole layout clues. Similar to the LayoutLM/LayoutLMv2, we train the LayoutXLM with the Multilingual Masked Visual-Language Modeling objective (MMVLM).

In LayoutLM/LayoutLMv2, an English word is treated as the basic unit, and its layout information is obtained by extracting the bounding box of each word with OCR tools, then subtokens of each word share the same layout information. However, for LayoutXLM, this strategy is not applicable because the definition of the linguistic unit is different from language to language. To prevent the language-specific pre-processing, we decide to obtain the character-level bounding boxes. After the tokenization using SentencePiece with a unigram language model, we calculate the bounding box of each token by merging the bounding boxes of all characters it contains. In this way, we can efficiently unify the multilingual multimodal inputs.

Text-Image Alignment The Text-Image Alignment (TIA) task is designed to help the model capture the fine-grained alignment relationship between text and image. We randomly select some text lines and then cover their corresponding image regions on the document image. The model needs to predict a binary label for each token based on whether it is covered or not.

Text-Image Matching For Text-Image Matching (TIM), we aim to align the high-level semantic representation between text and image. To this end, we require the model to predict whether the text and image come from the same document page.

3.3 Pre-training Data

The LayoutXLM model is pre-trained with documents in 53 languages. In this section, we briefly describe the pipeline for preparing the large-scale multilingual document collection.

Data Collection To collect a large-scale multilingual visually rich document collection, we download and process publicly available multilingual digital-born PDF documents following the principles and policies of Common Crawl[2]. Using digital-born PDF documents can benefit the collecting and pre-processing steps. On the one hand, we do not have to identify scanned documents among the natural images. On the other hand, we can directly extract accurate text with corresponding layout information with off-the-shelf PDF parsers and save time for running expensive OCR tools.

Pre-processing The pre-processing step is needed to clean the dataset since the raw multilingual PDFs are often noisy. We use an open-source PDF parser called PyMuPDF[3] to extract text, layout, and document images from PDF documents. After PDF parsing, we discard the documents with less than 200 characters. We use the language detector from the FastText (Joulin et al., 2017) library and split data per language. Following CCNet (Wenzek et al., 2020), we classify the document as the language if the language score is higher than 0.5. Otherwise, unclear PDF files with a language score of less than 0.5 are discarded.

Data Sampling After splitting the data per language, we use the same sampling probability $p_l \propto (n_l/n)^\alpha$ as XLM (Conneau and Lample, 2019) to sample the batches from different languages, where n_l is the document counts per language and n denotes the total number. Following InfoXLM (Chi et al., 2021), we use $\alpha = 0.7$ for LayoutXLM to make a reasonable compromise between performance on high- and low-resource languages. Finally, we follow this distribution and sample a multilingual document dataset with 22 million visually rich documents. In addition, we also sample 8 million scanned English documents from the IIT-CDIP dataset so that we totally use 30 million documents to pre-train the LayoutXLM, where the model can benefit from the visual information of both scanned and digital-born document images.

4 Key-value Extraction with PLMs

In this section, we present a simple yet efficient baseline framework based on pre-trained language

[2] https://commoncrawl.org
[3] https://github.com/pymupdf/PyMuPDF

models (PLMs) for our two sub-tasks. Equipped with this framework, we integrate two existing popular cross-lingual pre-trained language models, XLM-RoBERTa and InfoXLM, and our proposed LayoutXLM as the pre-trained language model backbones.

In this framework, given a visually rich document $\mathcal{D}$, we will pass discrete token set $\mathbf{T} = \{t_0, t_1, ..., t_n\}$ into these backbone models to obtain the contextual representation of each tokens $\mathbf{H} = \{\mathbf{h_0}, \mathbf{h_1}, ..., \mathbf{h_n}\}$. For different tasks, the representations will be processed with different modules to predict the required labels.

4.1 Semantic Entity Recognition

For this task, we simply follow the typical sequence labeling paradigm with BIO labeling format and build task-specific feed-forward network layers ($\texttt{FFN}^{SER}$) over the output of backbone models.

$$\mathbf{h}_i^{SER} = \texttt{FFN}^{SER}(\mathbf{h}_i)$$

4.2 Relation Extraction

For the relation extraction task, we first incrementally construct the set of relation candidates by producing all possible pairs of given semantic entities. For each pair, the representation of the head entity $\mathbf{h}_i^{head}$ or tail entity $\mathbf{h}_j^{tail}$ is the concatenation of the first token vector in each entity and the entity type embedding $\mathbf{e}^{head}$/$\mathbf{e}^{tail}$ obtained with a specific type embedding layer. After respectively projected by two feed-forward network layers, the representations of head and tail are fed into a biaffine classifier consisting of trainable weights $\mathbf{U}$, $\mathbf{W}$, and $\mathbf{b}$.

$$\begin{aligned}
\mathbf{h}_i^{head} &= \texttt{FFN}^{head}([\mathbf{h}_i; \mathbf{e}^{head}]) \\
\mathbf{h}_j^{tail} &= \texttt{FFN}^{tail}([\mathbf{h}_j; \mathbf{e}^{tail}]) \\
\mathbf{h}_{i,j}^{relation} &= \mathbf{h}_i^{head}\mathbf{U}\mathbf{h}_j^{tail} + \mathbf{W}(\mathbf{h}_i^{head} \circ \mathbf{h}_j^{tail}) + \mathbf{b}
\end{aligned}$$

5 Experiments

5.1 Settings

Cross-lingual Evaluation Besides the experiments of typical language-specific fine-tuning, we also design two additional settings to demonstrate the ability to transfer knowledge among different languages, which are zero-shot transfer learning and multitask fine-tuning. 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 trained on English data only and then evaluated on each target language. (3) Multitask fine-tuning requires the model to train on data in all languages. We evaluate models in these three settings over two sub-tasks in XFUND: semantic entity recognition and relation extraction, and compare LayoutXLM to two cross-lingual language models: XLM-R and InfoXLM.

Pre-training LayoutXLM Following the original LayoutLMv2 recipe, we train LayoutXLM models with two model sizes. For the $\text{LayoutXLM}_{\text{BASE}}$ model, we use a 12-layer Transformer encoder with 12 heads and set the hidden size to $d = 768$. For the $\text{LayoutXLM}_{\text{LARGE}}$ model, we increase the layer number to 24 with 16 heads and hidden size to $d = 1,024$. ResNeXt101-FPN is used as a visual backbone in both models. Finally, the number of parameters in these two models are approximately 345M and 625M. During the pre-training stage, we first initialize the Transformer encoder along with text embeddings from InfoXLM and initialize the visual embedding layer with a Mask-RCNN model trained on PubLayNet. The rest of the parameters are initialized randomly. Our models are trained with 64 Nvidia V100 GPUs with batch size of 1,024 for 150k training steps.

Fine-tuning on XFUND For a fair comparison, we train all models with the basic hyper-parameter settings and slightly adapt them to make sure every optimization has well converged. For the semantic entity recognition task, we train for 1,000 steps with batch size of 16. For the relation extraction task, we train for 3,000 steps with batch size of 8. We use the linear decay with a learning rate of 5e-5 and warm-up ratio of 0.1.

5.2 Results

We evaluate the LayoutXLM model on language-specific fine-tuning tasks, and the results are shown in Table 2. Compared with the pre-trained models such as XLM-R and InfoXLM, the LayoutXLM LARGE model achieves the highest F1 scores in both SER and RE tasks. The significant improvement shows LayoutXLM's capability to transfer knowledge obtained from pre-training to downstream tasks, which further confirms the effectiveness of our multilingual pre-training framework.

For the cross-lingual zero-shot transfer, we present the evaluation results in Table 3. Although the models are only fine-tuned on FUNSD dataset (in English), it can still transfer the knowledge to

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	XLM-RoBERTa$_{\text{BASE}}$	0.667	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$_{\text{BASE}}$	**0.794**	**0.8924**	**0.7921**	**0.7550**	**0.7902**	**0.8082**	**0.8222**	**0.7903**	**0.8056**
	XLM-RoBERTa$_{\text{LARGE}}$	0.7074	0.8925	0.7817	0.6515	0.7170	0.7139	0.711	0.7241	0.7374
	InfoXLM$_{\text{LARGE}}$	0.7325	0.8955	0.7904	0.6740	0.7140	0.7152	0.7338	0.7212	0.7471
	LayoutXLM$_{\text{LARGE}}$	**0.8225**	**0.9161**	**0.8033**	**0.7830**	**0.8098**	**0.8275**	**0.8361**	**0.8273**	**0.8282**
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$_{\text{BASE}}$	**0.5483**	**0.7073**	**0.6963**	**0.6896**	**0.6353**	**0.6415**	**0.6551**	**0.5718**	**0.6432**
	XLM-RoBERTa$_{\text{LARGE}}$	0.3473	0.6475	0.6798	0.6330	0.6080	0.6171	0.6189	0.5762	0.5910
	InfoXLM$_{\text{LARGE}}$	0.3679	0.6775	0.6604	0.6346	0.6096	0.6659	0.6057	0.5800	0.6002
	LayoutXLM$_{\text{LARGE}}$	**0.6404**	**0.7888**	**0.7255**	**0.7666**	**0.7102**	**0.7691**	**0.6843**	**0.6796**	**0.7206**

Table 2: Language-specific fine-tuning accuracy (F1) on the XFUND dataset (fine-tuning on X, testing on X), where "SER" denotes the semantic entity recognition and "RE" denotes the relation extraction.

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	XLM-RoBERTa$_{\text{BASE}}$	0.667	0.4144	0.3023	0.3055	0.371	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$_{\text{BASE}}$	**0.794**	**0.6019**	**0.4715**	**0.4565**	**0.5757**	**0.4846**	**0.5252**	**0.539**	**0.5561**
	XLM-RoBERTa$_{\text{LARGE}}$	0.7074	0.5205	0.3939	0.3627	0.4672	0.3398	0.418	0.4997	0.4637
	InfoXLM$_{\text{LARGE}}$	0.7325	0.5536	0.4132	0.3689	0.4909	0.3598	0.4363	0.5126	0.4835
	LayoutXLM$_{\text{LARGE}}$	**0.8225**	**0.6896**	**0.519**	**0.4976**	**0.6135**	**0.5517**	**0.5905**	**0.6077**	**0.6115**
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$_{\text{BASE}}$	**0.5483**	**0.4494**	**0.4408**	**0.4708**	**0.4416**	**0.4090**	**0.3820**	**0.3685**	**0.4388**
	XLM-RoBERTa$_{\text{LARGE}}$	0.3473	0.2421	0.3037	0.2843	0.2897	0.2496	0.2617	0.2333	0.2765
	InfoXLM$_{\text{LARGE}}$	0.3679	0.3156	0.3364	0.3185	0.3189	0.2720	0.2953	0.2554	0.3100
	LayoutXLM$_{\text{LARGE}}$	**0.6404**	**0.5531**	**0.5696**	**0.5780**	**0.5615**	**0.5184**	**0.4890**	**0.4795**	**0.5487**

Table 3: Zero-shot transfer accuracy (F1) on the XFUND dataset (fine-tuning on FUNSD, testing on X), where "SER" denotes the semantic entity recognition and "RE" denotes the relation extraction.

different languages. In addition, it is observed that the LayoutXLM model significantly outperforms the other text-based models. This verifies that LayoutXLM can capture the common layout invariance among languages and transfer to others.

Finally, Table 4 shows the evaluation results on the multitask learning. In this setting, the pre-trained LayoutXLM model is fine-tuned with all 8 languages simultaneously and evaluated on each specific language, in order to investigate whether improvements can be obtained by multilingual fine-tuning. We observe that the multitask learning further improves the model performance compared to the language-specific fine-tuning, which also confirms that document understanding can benefit from the layout invariance among different languages.

6 Related Work

Multimodal Pre-training Multimodal pre-training has become popular in recent years due to its successful applications in vision-language representation learning. Lu et al. (2019) proposed ViLBERT for learning task-agnostic joint representations of image content and natural language by extending the popular BERT architecture to a multimodal two-stream model. Su et al. (2020) proposed VL-BERT that adopts the Transformer model as the backbone, and extends it to take both visual and linguistic embedded features as input. Li et al. (2020a) propose VisualBERT consists of a stack of Transformer layers that implicitly align elements of an input text and regions in an associated input image with self-attention. Chen et al. (2020) introduced UNITER that learns through large-scale pre-training over four image-text datasets (COCO, Visual Genome, Conceptual Captions, and SBU Captions), which can power heterogeneous downstream V+L tasks with joint multimodal embeddings. Li et al. (2020b) proposed a new learning method Oscar (Object-Semantics Aligned Pre-training), which uses object tags

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	XLM-RoBERTa$_{\text{BASE}}$	0.6633	0.883	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$_{\text{BASE}}$	**0.7924**	**0.8973**	**0.7964**	**0.7798**	**0.8173**	**0.821**	**0.8322**	**0.8241**	**0.8201**
	XLM-RoBERTa$_{\text{LARGE}}$	0.7151	0.8967	0.7828	0.6615	0.7407	0.7165	0.7431	0.7449	0.7502
	InfoXLM$_{\text{LARGE}}$	0.7246	0.8919	0.7998	0.6702	0.7376	0.7180	0.7523	0.7332	0.7534
	LayoutXLM$_{\text{LARGE}}$	**0.8068**	**0.9155**	**0.8216**	**0.8055**	**0.8384**	**0.8372**	**0.853**	**0.8650**	**0.8429**
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$_{\text{BASE}}$	**0.6671**	**0.8241**	**0.8142**	**0.8104**	**0.8221**	**0.8310**	**0.7854**	**0.7044**	**0.7823**
	XLM-RoBERTa$_{\text{LARGE}}$	0.4246	0.7316	0.7350	0.7513	0.7532	0.7520	0.7111	0.6582	0.6896
	InfoXLM$_{\text{LARGE}}$	0.4543	0.7311	0.7510	0.7644	0.7549	0.7504	0.7356	0.6875	0.7037
	LayoutXLM$_{\text{LARGE}}$	**0.7683**	**0.9000**	**0.8621**	**0.8592**	**0.8669**	**0.8675**	**0.8263**	**0.8160**	**0.8458**

Table 4: Multitask fine-tuning accuracy (F1) on the XFUND dataset (fine-tuning on 8 languages all, testing on X), where "SER" denotes the semantic entity recognition and "RE" denotes the relation extraction.

detected in images as anchor points to significantly ease the learning of alignments. Inspired by these vision-language pre-trained models, we would like to introduce the vision-language pre-training into the document intelligence area, where the text, layout, and image information can be jointly learned to benefit the VRDU tasks.

Multilingual Pre-training Multilingual pre-trained models have pushed many SOTA results on cross-lingual natural language understanding tasks by pre-training the Transformer models on different languages. These models have successfully bridged the language barriers in many cross-lingual transfer benchmarks such as XNLI (Conneau et al., 2018) and XTREME (Hu et al., 2020). Devlin et al. (2019) introduced a new language representation model called BERT and extend to a multilingual version called mBERT, which is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create SOTA models for a wide range of tasks. Conneau and Lample (2019) proposed two methods to learn cross-lingual language models (XLMs): one unsupervised that only relies on monolingual data, and one supervised that leverages parallel data with a new cross-lingual language model objective. Conneau et al. (2020) proposed to train a Transformer-based masked language model on 100 languages, using more than two terabytes of filtered CommonCrawl data, which significantly outperforms mBERT on a variety of cross-lingual benchmarks. Recently, Chi et al. (2021) formulated cross-lingual language model pre-training as maximizing mutual information between multilingual-multi-granularity texts. The unified view helps to better understand the existing methods for learning cross-lingual representations, and the information-theoretic framework inspires to propose a pre-training task based on contrastive learning. Liu et al. (2020) presented mBART – a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective. Xue et al. (2021) introduced mT5, a multilingual variant of T5 that was pre-trained on a new Common Crawl-based dataset covering 101 languages. The pre-trained LayoutXLM model is built on the multilingual textual models as the initialization, which benefits the VRDU tasks in different languages worldwide.

7 Conclusion

In this paper, we introduce the multilingual form understanding benchmark XFUND, which includes key-value labeled forms in 7 languages. Meanwhile, we present LayoutXLM, a multimodal pre-trained model for multilingual visually rich document understanding. We make XFUND and LayoutXLM publicly available to advance the document understanding research. For future research, we will further enlarge the multilingual training data to cover more languages as well as more document layouts and templates. In addition, as there are a great number of business documents with the same content but in different languages, we will also investigate how to leverage the contrastive learning of parallel documents for the multilingual pre-training.

References

Haithem Afli and Andy Way. 2016. Integrating optical character recognition and machine translation of historical documents. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 109–116, Osaka, Japan. The COLING 2016 Organizing Committee.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020*, pages 104–120, Cham. Springer International Publishing.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. 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*, SIGIR '06, page 665–666, New York, NY, USA. Association for Computing Machinery.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020a. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision – ECCV 2020*, pages 121–137, Cham. Springer International Publishing.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

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*.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: Key information extraction datasets involving long documents with complex layouts. In *Document Analysis and Recognition – ICDAR 2021*, pages 564–579, Cham. Springer International Publishing.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1192–1200. ACM.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Ethical Consideration

The ethical implications of research are always an important consideration for us. While pursuing better model performance and high quality datasets, we respect the intellectual property rights of data resources, the privacy and rights of data sources, and strive to avoid potential harm to vulnerable populations.

When crawling the documents needed to build the XFUND dataset and LayoutXLM pre-training data, we strictly follow each site's robots exclusion standard [4] to ensure we are allowed to collect data. We also manually excluded websites with privacy concerns, keeping only those pages that we had permission to edit and republish according to the permission rules.

For the data used to build XFUND, we first removed all content and kept only the template, thus removing the maximum amount of sensitive content. On this basis, annotators filled in the templates using synthetic data that does not involve sensitive personal information of annotators, thus ensuring the privacy and rights of annotators. Then, we manually reviewed the templates to prevent potential privacy violations and harm to vulnerable populations. Any data that does not meet the specifications will be completely deleted.

B LayoutXLM

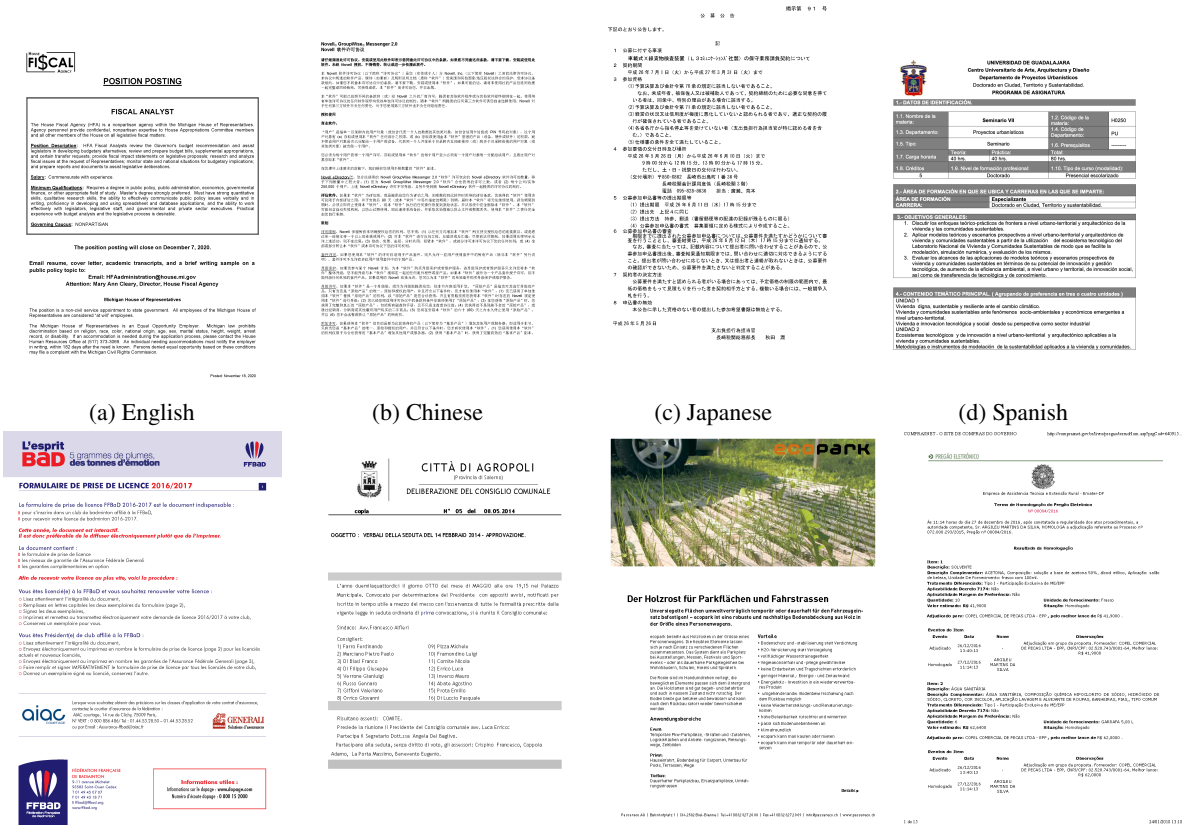
B.1 Pre-training Data Samples

We show pre-training samples of each languages in Figure 4.

B.2 Pre-training Data Distribution

Figure 5 shows the complete list of languages with the distribution of pre-training languages.

[4] https://en.wikipedia.org/wiki/Robots_exclusion_standard



(a) English (b) Chinese (c) Japanese (d) Spanish

(e) French (f) Italian (g) German (h) Portuguese

Figure 4: Real-world business documents with different layouts and languages for pre-training LayoutXLM

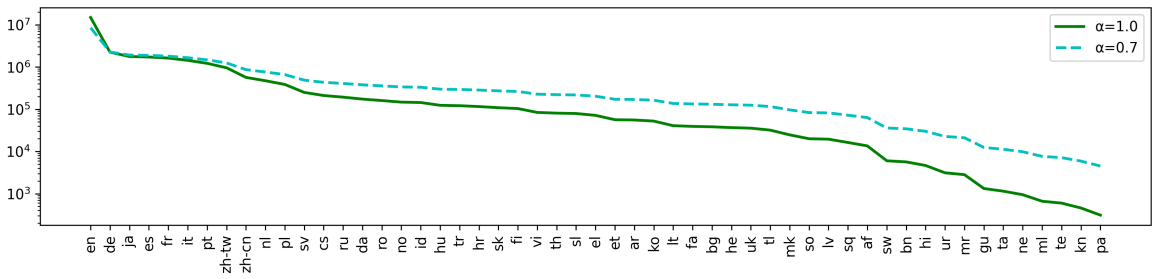


Figure 5: Language distribution of the data for pre-training LayoutXLM. We also show the document counts per language for different sampling exponents α.