Read Extensively, Focus Smartly: A Cross-document Semantic Enhancement Method for Visual Documents NER

Jun Zhao[1]*, **Xin Zhao**[1]*, **Wenyu Zhan**[1], **Tao Gui**[2]†, **Qi Zhang**[1]†,
Liang Qiao[3], **Zhanzhan Cheng**[3], **Shiliang Pu**[3]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai, China
[2]Institute of Modern Languages and Linguistics, Fudan University
[3]Hikvision Research Institute, Shanghai, China
{zhaoj19,tgui,qz}@fudan.edu.cn,{zhaoxin21,wyzhan21}@m.fudan.edu.cn

Abstract

The introduction of multimodal information and pretraining technique significantly improves entity recognition from visually-rich documents. However, most of the existing methods pay unnecessary attention to irrelevant regions of the current document while ignoring the potentially valuable information in related documents. To deal with this problem, this work proposes a cross-document semantic enhancement method, which consists of two modules: 1) To prevent distractions from irrelevant regions in the current document, we design a learnable attention mask mechanism, which is used to adaptively filter redundant information in the current document. 2) To further enrich the entity-related context, we propose a cross-document information awareness technique, which enables the model to collect more evidence across documents to assist in prediction. The experimental results on two documents understanding benchmarks covering eight languages demonstrate that our method outperforms the SOTA methods.

1 Introduction

Visually-rich documents (VRDs) are the most common information carriers in the real world, such as newspapers, resumes, tickets, etc. Different from plain text data, the information in VRDs is encoded by multiple modalities including text, vision, and layout. Entity recognition from VRDs, as a key step for document understanding, is of utmost practical interest for many downstream applications such as business analysis (Xu et al., 2020), intelligent education (Kahraman et al., 2010), digital library (Kroll et al., 2021).

The pioneering explorations approach this task by either computer vision (Katti et al., 2018) or natural language processing (Lample et al., 2016) paradigm. However, these methods ignore the inherent multi-modality of VRDs and consequently the suboptimal results are achieved. To address this problem, graph neural network (Liu et al., 2019) and self-attention mechanism (Zhang et al., 2020) are introduced to capture the cross-modality interaction and achieve superior performance. Recently, inspired by the widespread success of large pretrained language models (Qiu et al., 2020), self-supervised pretraining techniques are leveraged to learn cross-modal knowledge from unlabeled documents (Xu et al., 2020, 2021a,b) and the amount of labeled data required for document understanding is greatly reduced.

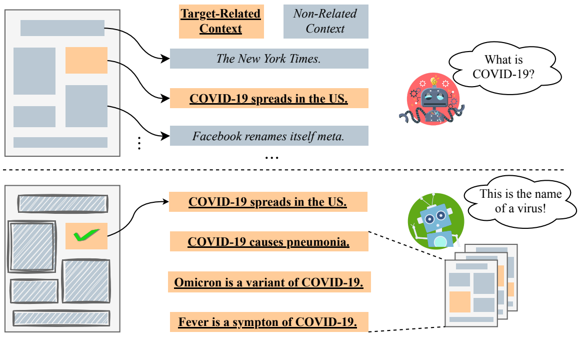
Target-Related Context
Non-Related Context
The New York Times.
COVID-19 spreads in the US.
Facebook renames itself meta.
...
What is COVID-19?
COVID-19 spreads in the US.
COVID-19 causes pneumonia.
Omicron is a variant of COVID-19.
Fever is a sympton of COVID-19.
This is the name of a virus!

Figure 1: Redundant irrelevant information and insufficient entity-related information in current document make it difficult to extract entities accurately. We filter redundant information from irrelevant regions and expand the information source from a single to multiple related documents. Thereby, entity-related information is efficiently integrated and the extraction is improved.

However, the existing methods suffer from the following two main shortcomings. From **intra-document** perspective, most previous works model a document by encoding each token uniformly, ignoring the fact that the document is composed of several regions (e.g. paragraphs, tables, captions) that are not so closely related semantically (Binmakhashen and Mahmoud, 2019). As shown in Figure 1, the type of COVID-19 in

*Equal Contributions.

†Corresponding authors.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2034–2043
October 12–17, 2022.

the orange region does not depend on the context in other gray regions. Unrestricted consideration of the whole document will not only distract the attention to local regions with stronger semantic associations (Guo et al., 2019b) but also increase the risk of fitting spurious features (also known as the *Shortcut* phenomenon (Geirhos et al., 2020)). From the **inter-document** perspective, the observed information is limited to the single document context, which may not be enough to accurately recognize entities. As shown in Figure 1, we can't judge whether COVID-19 is a piece of news or a virus only by virtue of "COVID-19 spreads in the US". Worse still, the insufficient context information can be further destroyed by the errors in character recognition when faced with low-quality documents (e.g. imaging blur, deformity) (Mou et al., 2020).

In this work, we propose a cross-document semantic enhancement method to enable the model to focus and integrate entity-related information across documents. Specifically, to prevent distractions from irrelevant regions in the current document, we design a learnable attention mask mechanism. Cross-region attentions are masked with a higher probability, so the model tends to make predictions using more reliable features from the local region. We introduce gumbel softmax (Jang et al., 2017) to solve the non-differentiable problem of discrete mask variables. To embrace sufficient context to assist prediction, we propose a cross-document information awareness technique. Inverted index (Knuth, 1997) is used to efficiently store and retrieve the contextualized representations of each token from each document. Cross-document attention acts on entity token representations to collect sufficient evidence to improve prediction.

Our contribution is threefold: 1) we propose a cross-document semantic enhancement method to enable the model to focus on the entity-related information across documents. The method filters the redundant information while expanding the information source; 2) our method can be regarded as a plug-in. It can be added to any document understanding model to improve prediction. 3) The experimental results on two datasets covering eight languages demonstrate that the proposed method outperforms the state-of-the-art methods;

2 Related Work

2.1 Multi-modal Named Entity Recognition

The approaches used to handle the task roughly fall into one of three directions. (1) **From single to multiple modality**. Due to the inherent multi-modality of visually-rich documents, early attempts from the perspective of computer vision (CV) (Katti et al., 2018) or natural language processing (NLP) (Lample et al., 2016; Ma et al., 2022; Zhao et al., 2021; Wang et al., 2022) can not achieve optimal performance. Therefore, artfully designed network architectures (Yu et al., 2021) and sophisticated mechanisms (Guo et al., 2019a) are used to fuse multimodal features for document understanding. (2) **From isolated to end-to-end optimization**. In classical document understanding methods, modules such as text recognition, image encoding, and information extraction are still optimized in isolation with different objective functions. To deal with this limitation and extract task-tailored features, end-to-end training frameworks are proposed (Zhang et al., 2020; Wang et al., 2021a), in which all modules are differentiable and optimized by a unified loss function. (3) **From supervised to pretraining paradigm**. Recently, Xu et al. (2020) extend the textual pretraining task to visual documents. Multiple well-designed pretraining objectives facilitate the interaction of multimodal information. With the help of learned generic document features, only a few samples would be sufficient to achieve SOTA accuracy. Different from all the above-mentioned methods, we rethink this task from the perspective of information integration and achieve the transition **from single-document uniform to cross-document selective information integration** to improve prediction.

2.2 Shortcut Phenomenon in Neural Network

Why should we selectively focus on local regions rather than dealing with each token of the whole document without distinction? The answer is the Shortcut phenomenon (Geirhos et al., 2020), which illustrates that neural networks always tend to fit training objectives in the simplest way. For example, suppose that in the training set, the word "Washington" in samples entitled "New York Times" all stand for place names. Global self-attention can easily learn this spurious feature. When "Washington" appears as a person's name, the model may completely ignore its context and

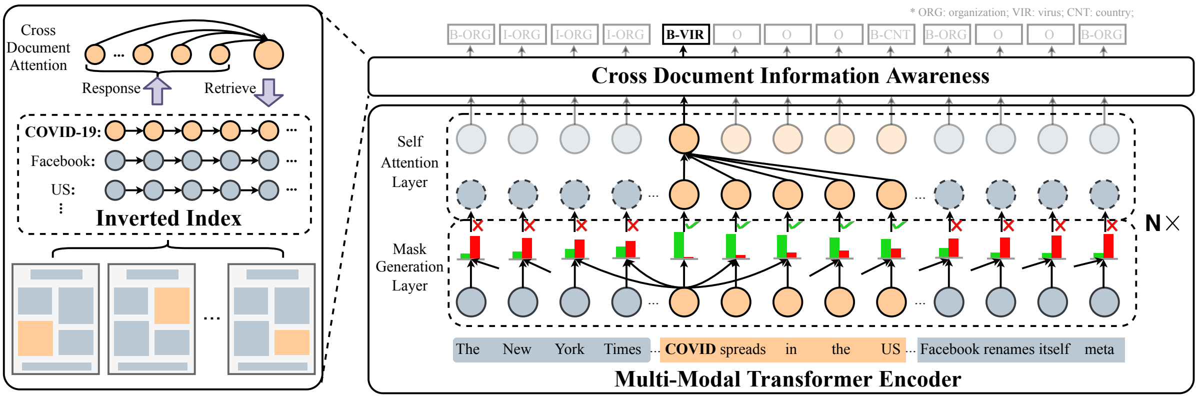

Figure 2: We take the extraction of COVID-19 as an example to illustrate the proposed method. First, a multimodal encoder is used to encode the context information of the current document. To filter the redundant irrelevant information (gray nodes) in the document, the mask generation layer is added to each layer of the encoder. Tokens farther away from COVID-19 will be masked with a higher probability during training (dotted node). Finally, we efficiently query COVID-19's representation in other documents by inverted index and a cross-document self-attention is used to integrate the information of COVID-19 in related documents.

directly make wrong predictions based on the title.

3 Approach

We propose a cross-document semantic enhancement method, which enables the model to make use of the cross-document context of the target entity and reduce the impact of noise from irrelevant regions of the current document.

The problem settings in this paper are formally stated as follows. Let $\mathcal{E} = \{e_i\}_{i=1,\dots,N_\mathcal{E}}$ denotes the predefined $N_\mathcal{E}$ entity types of interest and $\mathcal{C} = \{c_i\}_{i=1,\dots,N_c}$ denotes the entity label set derived by $\mathcal{E}$ according to the BIO scheme. Let $\mathcal{D} = \{(w_i, \boldsymbol{b}_i, v_i)\}_{i=1,\dots,N_\mathcal{D}}$ be a visually-rich document (VRDs), where w_i denotes token in the document. $\boldsymbol{b}_i = (x_i^0, y_i^0, x_i^1, y_i^1)$ denotes the bounding box (i.e. the position in the document) and v_i denotes image patch of w_i. Given a labeled set of VRDs $\mathcal{S} = \{\mathcal{D}_i\}_{i=1,\dots,N_\mathcal{S}}$, we target at learning a mapping function $\mathcal{F} : \mathcal{D} \rightarrow \{c_{w_i}\}_{i=1,\dots,N_\mathcal{D}}$. In other words, $\mathcal{F}$ assigns an entity label $c_{w_i} \in \mathcal{C}$ to each word w_i in the document to extract entities of interest.

3.1 Method Overview

We improve entity recognition from VRDs by the proposed cross-document semantic enhancement method, which enables the model to focus on the entity-related information across documents, rather than being distracted by irrelevant regions in the current document. As illustrated in Figure 2, the proposed method works as follows.

(1) We encode a visually-rich document $\mathcal{D} \in \mathcal{S}$ using a multi-modal transformer encoder implemented as the pretrained LayoutXLM (Xu et al., 2021b), which takes multi-modal information including text w_i, layout $\boldsymbol{b}_i$ and picture patch cut by $\boldsymbol{b}_i$ as input, and $\boldsymbol{h}_{i,j}$ denotes the output of layer j of the encoder. However, a document is usually composed of several regions that are not so closely related semantically. Consequently, unrestricted consideration of the whole document will not only distract the attention to local regions with stronger semantic associations but also can increase the risk of fitting spurious features.

(2) To prevent distractions from irrelevant regions in the current document, we design a learnable attention mask mechanism. Tokens that are farther away from the current token will be masked with a higher probability. Since mask operation is a discrete variable sampled from the binomial distribution, Gumbel softmax is introduced to realize end-to-end optimization of mask distribution. However, the current document context may not contain enough information to accurately classify the output of the encoder $\boldsymbol{h}_{i,N}$ to the true entity label c_{w_i}.

(3) To embrace sufficient context to improve prediction, we propose a cross-document information awareness technique. Each word w_i corresponds to a queue $\mathcal{Q}_i = \{\boldsymbol{h}_{i,N}^m\}_{m=1}^{|\mathcal{Q}|}$ that stores the contextualized representation of w_i in $|\mathcal{Q}|$ different documents. When encoding the current document $\mathcal{D}$, for each word $w_i \in \mathcal{D}$, we retrieve $\mathcal{Q}_i$ and obtain the final representation

$\boldsymbol{h}_i$ through the cross-document attention between $\boldsymbol{h}_{i,N}$ and each $\boldsymbol{h}_{i,N}^m \in \mathcal{Q}_i$. After that, $\boldsymbol{h}_{i,N}$ is updated to queue $\mathcal{Q}_i$. The lazy update ensures efficiency. Based on $\boldsymbol{h}_i$, we classify w_i into its corresponding entity label c_{w_i}.

3.2 Learnable Attention Mask Mechanism

In this section, we elaborate on the proposed learnable attention mask mechanism. Firstly, the pretrained encoder aims to integrate multi-modal inputs such as text, layout, and vision, and obtain the fixed-length representation of each token. To reduce the excessive attention to the irrelevant regions in the document during encoding, a mask sampled from the binomial distribution is applied to the original attention distribution, and the tokens farther away from the current position will be masked with a higher probability. Finally, the introduction of Gumbel relaxation solves the problem of non-differentiability of the discrete mask, which enables the model to learn the optimal mask distribution in an end-to-end manner.

3.2.1 Multi-Modal Transformer Encoder

The proposed cross-document semantic enhancement method is architecture-agnostic and can be added to any encoder architecture based on self attention mechanism. We adopt LayoutXLM (Xu et al., 2021b) as the implementation of our encoder $\boldsymbol{enc}(\cdot)$ because LayoutXLM is a multilingual encoder, which enables us to comprehensively demonstrate the effectiveness of the proposed method in different languages. We follow the way in LayoutXLM to generate input of the encoder. Specifically, given a visually-rich document $\mathcal{D} = \{(w_i, \boldsymbol{b}_i, v_i)\}_{i=1,\dots,N_{\mathcal{D}}}$, the multi-modal transformer encoder $\boldsymbol{enc}(\cdot)$ takes inputs from three different modalities, including text w_i, layout $\boldsymbol{b}_i$, and image patch v_i, which are mapped to text embedding $\boldsymbol{w}_i$, layout embedding $\boldsymbol{PE}_i$, and visual embedding $\boldsymbol{v}_i$, respectively. The text and visual embeddings are concatenated, then plus the layout embedding to get the input embedding $\boldsymbol{H}^0 = \{\boldsymbol{h_{1,0}}, \boldsymbol{...}, \boldsymbol{h_{N,0}}\} \in \mathcal{R}^{N \times d}$. Then, the intra-document self-attention transform $\boldsymbol{H}^0$ into the queries $\boldsymbol{Q}^0 \in \mathcal{R}^{N \times d}$, the keys $\boldsymbol{K}^0 \in \mathcal{R}^{N \times d}$, and the values $\boldsymbol{V}^0 \in \mathcal{R}^{N \times d}$. Finally, the output of the current layer is calculated as follows:

$$\boldsymbol{H}^{l+1} = ATT(\boldsymbol{Q}^l, \boldsymbol{K}^l)\boldsymbol{V}^l \tag{1}$$

$$ATT(\boldsymbol{Q}^l, \boldsymbol{K}^l) = \text{Softmax}(\frac{\boldsymbol{Q}^l \boldsymbol{K}^{l,\top}}{\sqrt{d}}). \tag{2}$$

The output of the last layer $\boldsymbol{H}^N$ is used as the input of the entity classifier. Although the global attention mechanism can model the interaction between multi-modal information, the redundant irrelevant information contained in the document reduces the attention of the model to the local regions with stronger semantic relevance, which leads to sub-optimal results.

3.2.2 Self-Attention with Mask Mechanism

In order to enhance the attention to local regions and reduce the risk of fitting spurious features, we carefully design a mask generation layer, which aims to select a more reliable subset from the input document as the basis for model prediction. Specifically, for the representation of each token $\boldsymbol{h}_{i,l} \in \boldsymbol{H}^l$ in layer l of encoder, the mask generation layer outputs a specific mask sequence $\boldsymbol{m} = (m_0, m_1, ..., m_N)$, $m_i \in \{0, 1\}$, where 0 and 1 here represent discard and select respectively. When calculating whether a contextual token will be discarded, we first calculate its distance ΔX on the horizontal axis and ΔY on the longitudinal axis with the current word. Then the mask is sampled from the following binomial distribution:

$$\begin{aligned} P_{\Delta X,Y}(m) = m * \pi_{\Delta X,Y} \\ + (1-m) * (1 - \pi_{\Delta X,Y}) \end{aligned} \tag{3}$$

$$\pi_{\Delta X,Y} = e^{-[\alpha(\frac{\Delta X}{X_{MAX}})+\beta(\frac{\Delta Y}{Y_{MAX}})]}, \tag{4}$$

where X_{MAX} and Y_{MAX} denote the maximum height and width of the document, respectively. α and β are the learnable parameters. In addition to relative positions, we also try to take region type into account. However, mainstream document understanding datasets lack region labels. Due to the domain gap, the pseudo region labels obtained by the existing layout analyzer are too noisy to use. It should be noted that equation 4 is easy to extend. We only need to add more terms to the exponential term to consider more influencing factors (e.g. when region labels are available)

We concatenate the mask sequences $\boldsymbol{m} \in \mathcal{R}^N$ corresponding to each token to get a mask matrix $\boldsymbol{M} \in \mathcal{R}^{N \times N}$, which is used to refine the original attention distribution of token i in layer l.

$$\boldsymbol{E} = \frac{\boldsymbol{QK}^\top}{\sqrt{d}} \tag{5}$$

$$ATT_i(\boldsymbol{M}, \boldsymbol{Q}, \boldsymbol{K}) = \frac{\boldsymbol{M}_i exp(\boldsymbol{E}_i)}{\sum_{i'=1}^{N} \boldsymbol{M}_{i'} exp(\boldsymbol{E}_{i'})}. \tag{6}$$

Based on the ATT obtained from equation 6, we calculate $\boldsymbol{h}_{i,l+1}$ according to equation 1. It should be **emphasized** that although tokens farther away from the current token will be masked with a higher probability, this does not mean that the model can never observe them. The purpose of the above mechanism is to make the model tend to make predictions using more reliable local features.

3.2.3 Gumbel Relaxation

To solve the non-differentiable problem of discrete mask operation, we introduce Gumbel softmax (Jang et al., 2017) to approximate the mask operation from the categorical distribution so that it can be optimized through backpropagation. Since we are dealing with a 2-class sampling problem, the original Gumbel softmax approximation is reduced to sigmoid-form as follows:

$$Gumbel(L) = \frac{exp((L + G_1)/\tau)}{exp((L + G_1)/\tau) + exp(G_2/\tau)}, \tag{7}$$

where $Gumbel(\cdot)$ is a continuous approximation of discrete mask m_i, L denotes $logits = log(\frac{p}{1-p})$ and p is calculated according to equation 3. G_1 and G_2 are two noises sampled from Gumbel distribution (Gumbel, 1954) independently. In addition, $\tau \in (0, +\infty)$ denotes the temperature parameter. With the decrease of τ, the approximate result $Gumbel(\cdot)$ will gradually tend to one hot. In the inference, we directly use $P_{\Delta X,Y}(m = 1) = 0.5$ as the threshold of whether to mask a context word to ensure the consistency of inference results.

3.3 Cross-Document Information Awareness

For some hard cases, the context of the current document may not contain enough entity-related information. To embrace sufficient context to assist prediction, we propose the cross-document information awareness technique. Inverted index is introduced to deal with the fast retrieval of massive document context and cross-document attention is used to integrate entity-related information from different documents.

3.3.1 Efficient Retrieval Supported by Inverted Index

Efficient storage and retrieval of a large number of documents is a prerequisite for cross-document information awareness. To enable the model to efficiently retrieve different contexts containing the current token, we introduce the inverted index (Knuth, 1997) to manage the data efficiently. Specifically, each word w_i corresponds to a queue $\mathcal{Q}_i = \{\boldsymbol{h}_{i,N}^m\}_{m=1}^{|\mathcal{Q}|}$ that stores the contextualized representation of w_i in $|\mathcal{Q}|$ different documents. When the model queries the context of a token in other documents (e.g. COVID-19), we only need to return the queue corresponding to COVID-19 instead of traversing the entire dataset. However, as the model is updated, the contextualized representations stored in the queue will gradually become obsolete. It is obviously inefficient or even impossible to update the whole inverted index after the training of each batch. In order to solve this problem, we use the lazy update to maintain the queue. That is, after the current document $\mathcal{D}$ is encoded, we only update the representation $\boldsymbol{h}_{i,N}$ of each $w_i \in \mathcal{D}$ to the queue $\mathcal{Q}_i$. Finally, does storing these vectors incur excessive memory overhead? In fact, most words do not need to store more than 3 cross-document copies due to the long-tail effect. Words that appear many times are usually stop words, and it is useless to store too many of their representations. Therefore, we limit the maximum queue length to avoid useless storage. Overall, the queue size accounts for only about 5% to 10% of the encoder parameters.

3.3.2 Cross-Document Attention

Given the multi-modal embedding $\boldsymbol{h}_{i,N}$ of each word w_i outputted by layer N of encoder $\boldsymbol{enc}(\cdot)$, we integrate the information of different documents through the cross-document attention mechanism to assist prediction. First, we query the context representation queue $\mathcal{Q}_i = \{\boldsymbol{h}_{i,N}^m\}_{m=1}^{|\mathcal{Q}|} \in \mathcal{R}^{|\mathcal{Q}|\times d}$ corresponding to token w_i from the inverted index. Then $\boldsymbol{h}_{i,N}$ and $\boldsymbol{Q}$ are concatenated to get $\boldsymbol{H}^c \in \mathcal{R}^{(|\mathcal{Q}|+1)\times d}$, the input of cross-document attention. Then we transform $\boldsymbol{H}^c$ into the keys $\boldsymbol{K}^c \in \mathcal{R}^{(|\mathcal{Q}|+1)\times d}$, and the values $\boldsymbol{V}^c \in \mathcal{R}^{(|\mathcal{Q}|+1)\times d}$. We only calculate the query $\boldsymbol{q}^c$ of $\boldsymbol{h}_{i,N}$ for the efficiency of calculation. We obtain the representation $\boldsymbol{h}_i$ of the word w_i by integrating cross document information as follows:

$$\boldsymbol{h}_i = ATT(\boldsymbol{q}^c, \boldsymbol{K}^c)\boldsymbol{V}^c \tag{8}$$

$$ATT(\boldsymbol{q}^c, \boldsymbol{K}^c) = \text{Softmax}(\frac{\boldsymbol{q}^c \boldsymbol{K}^{c,\top}}{\sqrt{d}}). \tag{9}$$

Finally, a linear classifier $\boldsymbol{\eta}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{N_c}$ optimized by cross entropy transforms $\boldsymbol{h}_i$ to its corresponding entity label $c_{w_i} \in \mathcal{C}$.

4 Experimental Setup

In this section, we describe the datasets for training and evaluating the proposed method. We also detail the baseline models for comparison. Finally, we clarify the implementation details and hyperparameter configuration of our method.

4.1 Datasets

The effectiveness of the proposed method is not limited to a particular language. We conducted experiments on two well-known document understanding datasets consisting of eight languages to show the universality of our method.

FUNSD (Jaume et al., 2019) is an English dataset for form understanding in noisy scanned documents. It consists of 199 manually labeled real documents from different fields such as marketing, advertising, and scientific reports. Each entity is annotated by a label (i.e., question, answer, header, or other) following the BIO schema, and a bounding box indicating the 2D position in the document. The dataset is split into 149 training forms and 50 testing forms.

XFUND (Xu et al., 2021b) is a multilingual form understanding dataset, which extends the FUNSD dataset to 7 other languages including Chinese, Japanese, Spanish, French, Italian, German, and Portuguese. Each language includes 199 forms, where the training set includes 149 forms and the test set includes 50 forms.

4.2 Comparison Methods

To evaluate the effectiveness of our method, we select the following SOTA multilingual NER models for comparison. The first two baseline only use textual modal as self-supervised signal, while multimodal pretraining is used in the last baseline and achieves SOTA in document understanding task.

XLM-RoBERTa (Conneau et al., 2020) is a Transformer-based masked language model pretrained on one hundred languages, with more than two terabytes of data.

InfoXLM (Chi et al., 2021) is a cross-lingual pretrained model based on an information-theoretic framework. It formulates pretraining as maximizing mutual information between multilingual multi-granularity texts.

LayoutXLM (Xu et al., 2021b) is a multimodal pretrained model, which takes the information of three modalities (text, layout, and image) as input. The carefully designed cross-modal alignment pretraining objectives improve the effectiveness of visually-rich document modeling.

4.3 Implementation Details

We use the AdamW (2019) as the optimizer, with a learning rate of $5e-5$ and batch size of 4 for all datasets. The length of the queue is selected among $\{5, 10, 15, 20\}$ and experiments show that 10 is the best choice. We initialize α and β to be 0.2 and 1.0 respectively. τ is selected among $\{0.15, 0.25, 0.35\}$ and we choose the best one. We use the base version for all the pretrained models. All experiments are conducted using an NVIDIA GeForce RTX 3090 with 24GB memory. All experimental results are the average of three runs based on the Pytorch framework.

5 Results and Analysis

In this section, we present the experimental results on two well-known document understanding datasets to show the effectiveness of our method.

5.1 Main Results

Table 1 reports model performance on FUNSD, and XFUND datasets, which shows that the proposed method achieves state-of-the-art results in eight different languages. For visually-rich document understanding, the key information is presented in multiple modalities, such as text, layout, vision. However, XLM-RoBERTa and InfoXLM only model a single textual modal, consequently underperforming the multi-modal LayoutXLM baseline and our method by a large margin. Benefitting from (1) the irrelevant redundant information filtering supported by attention mask mechanism and (2) valuable entity-related context provided by the cross-document information awareness technique, the model can efficiently integrate entity-related information to make predictions. As a result, the proposed method outperforms LayoutXLM in eight different languages. This also shows that the effectiveness of the proposed semantic enhancement method is not limited to a specific language.

Dataset	Language	XLM-RoBERTa (Conneau et al., 2020)			InfoXLM (Chi et al., 2021)			LayoutXLM (Xu et al., 2021b)			**Ours**		
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
FUNSD	English	-	-	0.667	-	-	0.685	0.784	0.817	0.800	0.800	0.828	**0.814**
XFUND	Chinese	-	-	0.877	-	-	0.887	0.848	0.920	0.882	0.871	0.932	**0.901**
	Japanese	-	-	0.776	-	-	0.787	0.733	0.854	0.789	0.760	0.851	**0.803**
	Spanish	-	-	0.611	-	-	0.623	0.713	0.752	0.732	0.736	0.760	**0.748**
	French	-	-	0.674	-	-	0.702	0.762	0.794	0.778	0.778	0.807	**0.792**
	Italian	-	-	0.669	-	-	0.675	0.774	0.833	0.803	0.791	0.852	**0.820**
	German	-	-	0.681	-	-	0.706	0.771	0.824	0.797	0.783	0.836	**0.809**
	Portuguese	-	-	0.682	-	-	0.701	0.759	0.803	0.780	0.767	0.816	**0.790**
Avg.	ALL	-	-	0.705	-	-	0.721	0.768	0.825	0.795	**0.786**	**0.835**	**0.810**

Table 1: Main results on two well-known document understanding datasets.

Language	w/o. CD			w/o. AM			**Ours**		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
English	0.788	0.825	0.806	0.785	0.827	0.805	0.800	0.828	**0.814**
Chinese	0.868	0.934	0.900	0.869	0.928	0.897	0.871	0.932	**0.901**
Japanese	0.743	0.847	0.792	0.749	0.850	0.797	0.760	0.851	**0.803**
Spanish	0.718	0.753	0.735	0.717	0.755	0.736	0.736	0.760	**0.748**
French	0.763	0.802	0.782	0.767	0.811	0.789	0.778	0.807	**0.792**
Italian	0.785	0.844	0.814	0.776	0.846	0.809	0.791	0.852	**0.820**
German	0.779	0.820	0.799	0.766	0.826	0.795	0.783	0.836	**0.809**
Portuguese	0.757	0.809	0.782	0.766	0.808	0.787	0.767	0.816	**0.790**
Average	0.775	0.829	0.801	0.774	0.831	0.802	**0.786**	**0.835**	**0.810**

Table 2: Abalation study of our method.

5.2 Ablation Study

To study the contribution of each component in the proposed method, we conduct ablation experiments on the two datasets and display the results in Table 2. The results show that the model performance is degraded if the learnable attention mask (AM) is removed, indicating that the redundant information in the document will distract the model from focusing on the local regions with stronger semantic relevance. In addition, cross-document information awareness (CD) provides diverse contexts for key information extraction. Without CD, some hard cases where the entity-related information is insufficient can not be recognized accurately. Consequently, the overall performance will be negatively affected. It is worth noting that the proposed AM and CD are effective in all languages, which also shows the generality and practical value of the proposed method.

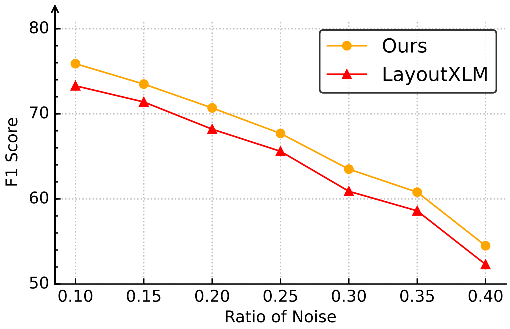

Figure 3: Model performance on the perturbated FUNSD dataset.

5.3 Robustness Analysis

In real-world applications, we inevitably encounter low-quality input documents. Due to various factors such as occlusion, focus and angular deformations, optical character recognition (OCR) of those documents often yield unsatisfactory results, which will negatively impact the information extraction. To exhaustively evaluate the robustness of the model in real scenarios, we apply TextFlint (Wang et al., 2021b), a robustness evaluation platform, to perturb the original dataset to simulate OCR errors in real-world applications. Specifically, we use OCR error transformation in TextFlint

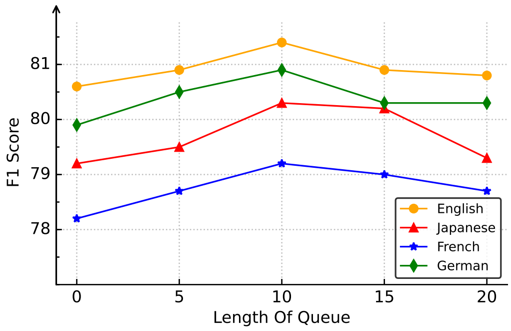

Figure 4: Model performance with different lengths of the queue in 4 randomly selected languages.

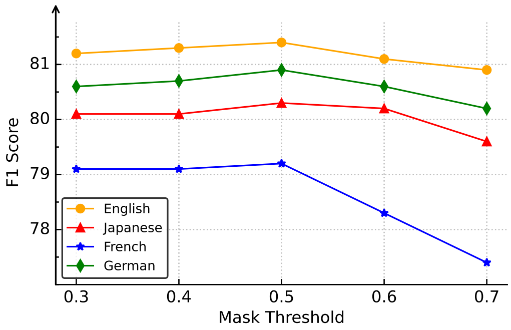

Figure 5: Model performance with different mask thresholds in 4 randomly selected languages.

to disturb the document. For each entity in the document, we perturb a certain proportion of words in its context and evaluate the prediction results. From Figure 3 we can see that the proposed method consistently outperforms baselines under the different ratios of OCR errors in context. Benefitting from the proposed cross-document information awareness technique, even if the context of entities in the current document is severely damaged, the model can still use the context information in related documents to assist in prediction. Therefore, the proposed model has better robustness in real scenarios.

5.4 Efficiency Analysis

Although the introduction of inverted index enables us to retrieve large-scale related documents efficiently, *a question worthy of discussion is whether storing the representations of each token will occupy too much memory space?* We answer the question by analyzing the maximum queue length L, a key parameter affecting memory consumption. From Figure 4 we can observe that optimal performance can be achieved by memorizing no more than 10 token representations of each target entity in related documents. In addition, nearly 70% of tokens in the two datasets appear no more than 3 times due to the long-tail token distribution. Therefore, the storage of cross-document information does not bring too high a memory overhead. In fact, too much information from other documents is not necessarily leading to better results. When the maximum queue length L exceeds 10, further increasing L will make a large number of meaningless stop words and some other outdated representations memorized, which leads to the decline of model performance.

5.5 Impact of Context Sparsity

In the inference, we directly use $P_{\Delta X,Y}(m = 1) = \theta$ as the threshold of whether to mask a context word to ensure the consistency of inference results. The threshold θ corresponds to the sparsity of the context that the model can observe. The larger the θ, the fewer tokens can be observed. To analyze the impact of context sparsity, we conduct experiments on four randomly selected languages. As can be seen from Fig. 5, initially, as the threshold increases, redundant noise from different regional contexts is continuously masked. The performance continues to improve, reaching the maximum when $\theta = 0.5$. After that, further increasing θ will cause more useful contexts in the same region to be masked, and the performance declines rapidly. This also confirms the view that useful context information is mainly distributed in local regions.

6 Conclusions

In this work, we introduce a cross-document semantic enhancement method to improve entity recognition from visually-rich documents. The proposed learnable attention mask mechanism effectively filters redundant irrelevant information in the current document, which reduces the risk of overfitting spurious features. Cross-document information awareness enriches sufficient entity-related context to improve predictions. The proposed method can be regarded as a plug-in, which can be added to any existing document understanding model and improve prediction. Experimental results show that the proposed method outperforms the existing state-of-the-art methods in documents of different languages and is more robust in real scenarios.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This research was sponsored by Hikvision Cooperation Fund.

References

Galal M Binmakhashen and Sabri A Mahmoud. 2019. Document layout analysis: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36.

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.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Emil Julius Gumbel. 1954. Statistical theory of extreme values and some practical applications : A series of lectures.

He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2019a. Eaten: Entity-aware attention for single shot visual text extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 254–259. IEEE.

Maosheng Guo, Yu Zhang, and Ting Liu. 2019b. Gaussian transformer: A lightweight approach for natural language inference. In *National Conference on Artificial Intelligence*.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

G. Jaume, H. Kemal Ekenel, and J. Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6, Los Alamitos, CA, USA. IEEE Computer Society.

H. Tolga Kahraman, Seref Sagiroglu, and Ilhami Colak. 2010. Development of adaptive and intelligent web-based educational systems. In *2010 4th International Conference on Application of Information and Communication Technologies*, pages 1–5.

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

Donald E. Knuth. 1997. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*, third edition. Addison-Wesley, Reading, Mass.

Hermann Kroll, Jan Pirklbauer, Jan-Christoph Kalo, Morris Kunz, Johannes Ruthmann, and Wolf-Tilo Balke. 2021. Narrative query graphs for entity-interaction-aware document retrieval. In *Towards Open and Trustworthy Digital Societies*, pages 80–95, Cham. Springer International Publishing.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 32–39, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ruotian Ma, Yiding Tan, Xin Zhou, Xuanting Chen, Di Liang, Sirui Wang, Wei Wu, and Tao Gui. 2022. Searching for optimal subword tokenization in cross-domain ner. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4289–4295. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yongqiang Mou, Lei Tan, Hui Yang, Jingying Chen, Leyuan Liu, Rui Yan, and Yaohong Huang. 2020. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 158–174. Springer.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271.

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

Xiao Wang, Shihan Dou, Limao Xiong, Yicheng Zou, Qi Zhang, Tao Gui, Liang Qiao, Zhanzhan Cheng, and Xuanjing Huang. 2022. MINER: Improving out-of-vocabulary named entity recognition from an information theoretic perspective. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5590–5600, Dublin, Ireland. Association for Computational Linguistics.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021b. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021a. 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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florêncio, Cha Zhang, and Furu Wei. 2021b. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *CoRR*, abs/2104.08836.

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 *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370. IEEE.

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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1413–1422.

Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. 2021. A relation-oriented clustering method for open relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9707–9718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.