KnowVrDU: A Unified Knowledge-aware Prompt-Tuning Framework for Visually-rich Document Understanding

Yunqi Zhang[1,2], Yubo Chen[3], jingzhe zhu[2], Jinyu Xu[2], Shuai Yang[5], Zhaoliang Wu[5], Liang Huang[5], Yongfeng Huang[1,3,4] and Shuai Chen[2]
[1] Department of Electronic Engineering & BNRist, Tsinghua University, Beijing, China
[2] Ant Group, Hangzhou, Zhejiang, China [3] Zhongguancun Laboratory, Beijing, China
[4] Institute for PrecisionMedicine, Tsinghua University, Beijing, China
[5] Health International Inc, Beijing, China
Corresponding author: yfhuang@mail.tsinghua.edu.cn

Abstract

In Visually-rich Document Understanding (VrDU), recent advances of incorporating layout and image features into the pre-training language models have achieved significant progress. Existing methods usually developed complicated dedicated architectures based on pre-trained models and fine-tuned them with costly high-quality data to eliminate the inconsistency of knowledge distribution between the pre-training task and specialized downstream tasks. However, due to their huge data demands, these methods are not suitable for few-shot settings, which are essential for quick applications with limited resources but few previous works are presented. To solve these problems, we propose a unified Knowledge-aware prompt-tuning framework for Visual-rich Document Understanding (KnowVrDU) to enable broad utilization for diverse concrete applications and reduce data requirements. To model heterogeneous VrDU structures without designing task-specific architectures, we propose to reformulate various VrDU tasks into a single question-answering format with task-specific prompts and train the pre-trained model with the parameter-efficient prompt tuning method. To bridge the knowledge gap between the pre-training task and specialized VrDU tasks without additional annotations, we propose a prompt knowledge integration mechanism to leverage external open-source knowledge bases. We conduct experiments on several benchmark datasets in few-shot settings and the results validate the effectiveness of our method.

Keywords: Visually-rich Document Understanding, Prompt Tuning, Knowledge Injection

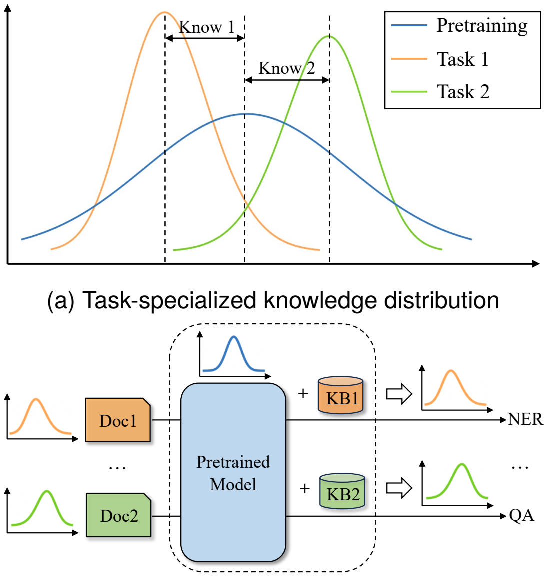
1. Introduction

Visually-rich Document Understanding (VrDU) seeks to automatically analyze and extract significant factual texts from image or digital-born documents, which is of immense academic and commercial value. VrDU encompasses extensive text-centric and image-centric tasks, notably document information extraction, document visual question answering, and document image classification, etc (Cui et al., 2021).

In recent years, pre-training techniques have been widely leveraged in the VrDU tasks to learn the multi-modal interactions between text, image and layout modalities. For example, Xu et al. (2020b) first introduced spatial layout information with 2-D position embeddings and integrated image embeddings into original BERT architecture. Peng et al. (2022) proposed to enhance the layout knowledge through the arrangement of reading order sequence. Huang et al. (2022) mitigated the discrepancy between text and image multi-modal representation learning with unified discrete token reconstructive objectives.

Existing methods achieved considerable success in capturing cross-modality information from digital documents. However, they naturally relied on huge demands of designing dedicated architectures and annotating task-specific samples to fine-tune pre-trained models. This is because the general knowledge of pre-trained models cannot completely encompass sophisticated distinct expertise of various specialized tasks, resulting in significant knowledge distribution gaps between the pre-training task and VrDU tasks, as shown in Figure 1 (a). Unfortunately, the data requirements make these methods typically expensive and difficult to achieve quick applications in practice (Wang and Shang, 2022), especially in few-shot scenarios with limited annotated data. Furthermore, under few-shot settings, limited annotations are insufficient and ineffective in capturing refined task expertise, thus leading to severe performance degradation on downstream VrDU tasks.

Our work is motivated by several observations. First, recent research of natural language processing uniformly model various tasks, including information extraction (Li et al., 2019; Zhou et al., 2022; Liu et al., 2022), document classification, and etc., with Question Answering (QA) format, such as extracting named entities from the text by asking what are the entities and answering the entity spans. These methods used only one QA model and distinguished different tasks with different question and answer templates, avoiding task-specific model design, which can be useful for VrDU. Sec-



(a) Task-specialized knowledge distribution

(b) A unified knowledge integration framework

Figure 1: (a) illustrates knowledge distribution gaps of the pre-training task and two specialized tasks. (b) shows our unified knowledge integration framework to bridge the knowledge gaps.

ond, the lack of task-specialized knowledge caused by limited labeled data in few-shot scenarios can be compensated with large-scale knowledge bases. For example, Wikipedia as a knowledge base can provide rich knowledge about rare named entities, which are almost unlikely to occur in limited annotated data like Attapeu (province in Laos) and Yakitori (Japanese skewered chicken), and enhance the performance of named entity recognition.

Based on the above observations, we propose a novel framework to uniformly model various VrDU tasks in few-shot scenarios. First, we propose to convert different VrDU tasks into a single QA format through task-specific prompts. Concretely, we adopt structural prompts (Zhong et al., 2022) to transform the input document image and its Optical Character Recognized (OCR) texts into template-generated question queries, where different templates are applied for different tasks. We also introduce continuous trainable prompt embeddings (Liu et al., 2021) as storable task adapters to make the model more flexible, and freeze the pre-trained model to enable parameter-efficient prompt tuning. Then, we propose a prompt knowledge injection mechanism to reduce the knowledge distribution gap between the general pre-training task and downstream VrDU tasks. We incorporate open-source knowledge bases for adequate task expertise, as shown in Figure 1 (b), and select informative knowledge through attention mechanism. We evaluate our Knowledge-aware prompt-tuning framework for VrDU (KnowVrDU) with three tasks: document information extraction, document classification and document question answering in few-shot settings.

The main contributions of this paper are:

- We propose a unified KnowVrDU framework to enable broad utilization for diverse concrete applications.
- To model heterogeneous VrDU structures without designing dedicated architectures, we propose to transform various VrDU tasks into a uniform question answering format with task-specific prompts and efficiently train the model through prompt tuning.
- To bridge the knowledge gap between pre-training and downstream tasks without additional annotations, we propose a prompt knowledge integration mechanism to incorporate external open-source knowledge bases.
- Experimental results on several benchmark datasets demonstrate the effectiveness of our method.

2. Related Work

In the early work of document intelligence, VrDU tasks were solved in feature-based approaches (O'Gorman, 1993; Simon et al., 1997; Shilman et al., 2005; Wei et al., 2013). For example, O'Gorman (1993) proposed document spectrum algorithm to analysis structural pages based on bottom-up clustering of page components. Wei et al. (2013) introduced statistical probabilistic models to learn the pixel features and detect physical structure of documents. However, these approaches were heavily restricted by hand-crafted expert knowledge and insufficient supervised data.

In recent years, neural network-based approaches became the dominant paradigm of visual document understanding, which focused on designing suitable network architectures. For example, Katti et al. (2018) represented the spatial structure of the document as a sparse 2D grid of characters and adopted CNN to perform semantic segmentation on 2D grids. Liu et al. (2019a) utilized graph convolutional networks to integrate the textual semantic information, layout of documents and relative positions of the individual segment. However, these approaches relied on sufficient supervised data and failed to model the joint representation of visual, text and layout features.

To address these issues, more recent work introduced pre-train-and-fine-tune paradigm to document intelligence, which focused on designing objectives used at both the pre-train and fine-tune stages. Xu et al. (2020b) first pre-trained BERT (Devlin et al., 2018) with masked visual-language model and multi-label document classification objectives to jointly model interactions between multi-modal information. Xu et al. (2020a) and Huang et al. (2022) further improved the pre-training strategies to strengthen the alignment among different modalities. Moreover, Zhao et al. (2022) proposed a cross-document semantic integration method to collect more evidence across documents in visual document NER. Besides, UDOP (Tang et al., 2023) unified pretraining and multi-domain downstream tasks into sequence generation scheme and pre-trained on both large-scale unlabeled and labeled data. However, the knowlegde gap between pre-train and fine-tune process hinders the expression of general knowledge in pre-trained models, resulting in the failure of existing models in few-shot scenarios. Though Wang and Shang (2022) proposed to embed the label surface names for few-shot entity recognition of document images, this method is not suitable to other VrDU tasks such as document question answering.

Different from previous work, we propose a unified framework for a wide applications of numerous VrDU tasks in few-shot scenarios. We propose to reformulate multiple VrDU tasks into a uniform question answering format. We also propose to leverage external open-source knowledge bases to bridge the knowledge gap without additional annotations. Experimental results in few-shot settings demonstrate the effectiveness of our method on a wide variety of downstream VrDU tasks.

3. Our Approach

3.1. Primilaries

The pre-train and fine-tune paradigm has already been proved to promisingly successful in Visually-rich Document Understanding. In this paradigm, multi-modal transformers are pre-trained on large-scale scanned document image datasets and fine-tuned with additional downstream architectures on specific tasks. However, in the few-shot scenarios, inadequate samples could not bridge the knowledge gap between pre-train and fine-tune stages, resulting in the inability to capture the specialized representations in other tasks. To tackle this challenge, we first propose a series of structural prompts to convert varying types of document understanding tasks to multi-span question answering. The structural prompts consist of fixed hard prompts to indicate the specific task information as well as tunable soft prompts to adjust with multi-modal language models. Then, we employ the knowledge attention network to enhance the representations of input data. The overall framework of our approach is shown in Figure 2.

3.2. Unified Structure Generation for VrDU

Our knowledge-aware prompt-tuning framework can be decomposed into two atomic operations: uniform prompts generation and external knowledge injection. This section describes how to transform heterogeneous VrDU structures into a formatted input to the model.

3.2.1. Overall Prompt Structure

The formatted prompts are constructed with multiple key-value pairs. Each key locates the relevant spans concerning to pre-defined description and each value contains the corresponding mentions. Concretely, we adopt four fixed keys in our prompts, including "*Task*", "*Question*", "*Passage*" and "*Knowledge*", which are placed sequentially in our structural prompts. Each key is surrounded by square brackets and separated from the value by a colon symbol. We introduce several soft prompts (Liu et al., 2021) which are tunable and storable as the values of "*Task*" indicator. There are two advantages of utilizing these trainable soft prompts. First of all, trainable embedddings enable us to find a better continuous prompts beyond the original vocabulary that the pre-trained model could express to make the model adaptively accommodate the structural prompts. Second, our model can discriminate different input components by these flexible prompts embedding and model the speciality of each task via task-specific values. Considering the differences in downstream tasks, distinct templates are leveraged to construct values of "*Question*". Furthermore, we use the original OCR results and visual features to initialize the values related to "*Passage*" key, as shown in Figure 2. Finally, we use "[*CLS*] [*Task*]: <*Soft Tokens*>, [*Question*]: <*Question Templates*>, [*Passage*]: <*Image*> <*OCR Tokens*>, [*Knowledge*]:<*Knowledge Contents*> [*SEP*]" as our final model input, as shown in Figure 2.

3.2.2. Downstream Task Transformation

We mainly focus on three downstream tasks: document information extraction, document visual question answering and document image classification. We modify the "<*Question Templates*>" to convert them to (multi-span) extractive QA tasks. Extractive QA involves identifying and recognizing (multiple) answer spans from a given passage. Given a question Q and the corresponding passage P, the model is required to extract the answer spans S of P and each instance in the QA dataset can be defined as (Q, P, S). Therefore, our aim is to convert inputs of heterogeneous structures into the form of ques-

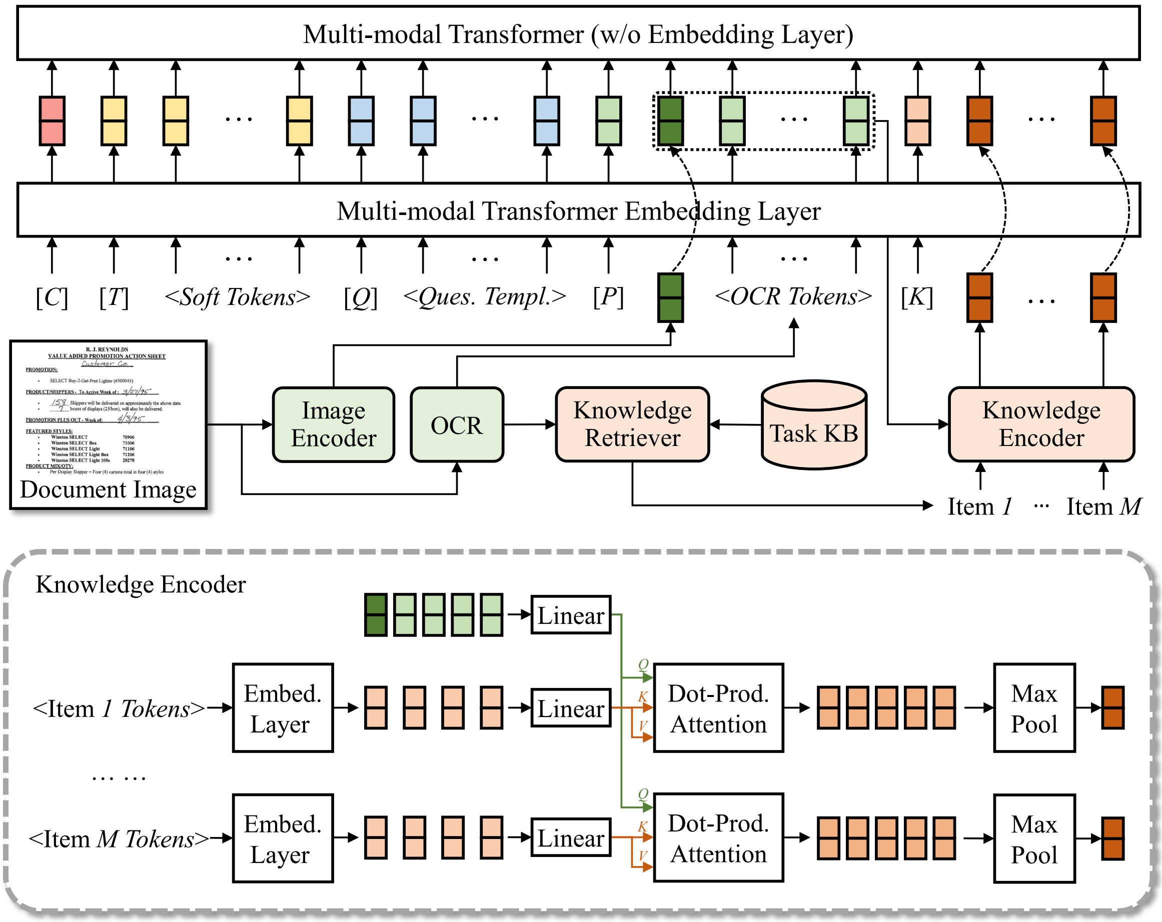


Figure 2: The overall framework of our approach. We decompose our knowledge-aware prompt-tuning framework into two atomic operations. We illustrate the uniform prompts generation in the upper part and the external knowl- edge attention injection process in the lower half part.

tion queries with assigned contexts from "*Passage*" values.

For **document information extraction**, we reformulate the original tasks as span-based extraction problem. Our model locates the start and end positions of the entity span by predicting the start/end scores of each tokens, which is essentially the same as original document named entity recognition task. Because each pair of start/end classifiers could only extract one type of entity, we adopt m templates for m entity types. As Liu et al. (2022) suggests, we choose the question templates like "*What is the* $[\mathrm{E}]$ *?*", where "$[\mathrm{E}]$ "is a placeholder which can be replaced by the name of any entity type. For example, in the case of the FUNSD (Jaume et al., 2019) dataset, "*What is the Document Header ?*" could be used to identity the "*Header*" entity type defined in the dataset.

For **document image classification**, we first transform the classification tasks into the extractive QA tasks by adopting the question template "*What type is the Document or Receipt ? <Question Options>* ". The question options are initialized with document labels and each a letter followed by a colon. Then we spot the start/end positions of document labels in our question template as new labels for the extractive QA task. For example, the question template utilized in categorizing the documents in RVI-CDIP (Rawat and Wang, 2017) could be "*What type is the Document or Receipt ? A: letter, B:memo, ...*".

For **document visual question answering**, we directly introduce the extractive QA tasks and use the original questions and passages without any modification.

In summary, our proposed prompts can effectively model different VrDU tasks as a uniform extractive QA task and discriminate different components with the four key indicators. Besides, the unified prompt structure provides the basis for task-specific knowledge injection in a unified form.

3.3. Knowledge Integration

In this section, we introduce our knowledge integration mechanism in detail, which begins with describing the representations of our structural prompts.

Then we propose the knowledge attention mechanism to incorporate external knowledge based into our prompts.

3.3.1. Prompt Representation

We utilize a multi-modal transformer to encode our structural prompts and learn cross-modal interactions. Given a structural prompt instance, we denote the representations of each key-value pair as $\mathbf{P}_i$, where $i \in [1, 2, 3, 4]$. Concretely, the key-value pair can further be represented as $\mathbf{P}_{\mathrm{i}} = \mathrm{Embedding}([D_i; V_i])$. D_i is the i-dx text description of key indicator and V_i is the i-dx corresponding value. Since we introduce the soft prompts as the value of $\mathbf{P}_1$ and $\mathbf{P}_4$, the pseudo tokens are adopted as placeholders to initialize the trainable embedding:

$$V_1 = [e_1, \ldots, e_n], V_4 = [e_1, \ldots, e_m] \tag{1}$$

where n and m are the length of pseudo tokens in "*Task*" and "*Knowledge*" values. Besides, V_2 stands for the tokens in question templates and V_3 is the concatenation of original text tokens and image tokens sequence. Finally, the final model input $\mathbf{E}_{\mathrm{P}}$ can be formed through concatenating all the key-value pairs $\mathbf{P}_i$:

$$\mathbf{E}_{\mathrm{P}} = \mathrm{Embedding}([D_1;V_1;D_2;V_2;D_3;V_3;D_4;V_4]) \tag{2}$$

Note that during tuning process we fix the parameters of key indicator $\mathrm{Embedding}(D_i)$ and the hard tokens embedding $\mathrm{Embedding}(V_2)$ and $\mathrm{Embedding}(V_3)$. We only train the soft prompts embedding $\mathrm{Embedding}(V_1)$ and $\mathrm{Embedding}(V_4)$ to learn the semantics of the structural prompts. Then the soft prompts can be saved to store the customized task-specific characteristics.

3.3.2. Knowledge Integration

Based on the OCR results, we employ the entity linker to perform entity linking with external worldwide knowledge base. The entity linking toolkit works by matching potential entity mentions in sentences to entity aliases from the knowledge base. We identify the semantic category for each detected entity and extract the entity description as supplementary expressions which usually contain the necessary task-specialized knowledge. These knowledge items help the model understand the given question and passage better.

However, integrating the external knowledge with the prompts immediately suffers from abundant noise of excessive irrelevant contents. Therefore, we propose to eliminate the noise and to refine task-specialized knowledge with the attention mechanism. We denote N as the length of retrieved knowledge contents and d_h as the dimension of transformer hidden states, the representation of extracted knowledge as $\mathbf{E}_{\mathrm{K}} \in \mathbb{R}^{N \times d_h}$. We adopt the dot-product attention (Vaswani et al., 2017a) to learn the fused representations of knowledge. Let the representation of OCR results be $\mathbf{E}_{\mathrm{T}} \in \mathbb{R}^{L \times d}$, we compute the attention query, key and value matrix:

$$\mathbf{Q} = \mathbf{E}_{\mathrm{T}}\mathbf{W}^Q, \mathbf{K} = \mathbf{E}_{\mathrm{K}}\mathbf{W}^V, \mathbf{V} = \mathbf{E}_{\mathrm{K}}\mathbf{W}^V \tag{3}$$

The $\mathbf{W}^Q, \mathbf{W}^V \in \mathbb{R}^{d_h \times d_k}$ are the trainable weight matrices. Then we introduce our attention layer as:

$$\mathbf{H} = \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

We perform the max-pooling and linear transformation operations $\mathbf{H}^{'} = \mathrm{MaxPooling}(\mathrm{Linear}(\mathbf{H}))$ and substitute the embedding of pseudo tokens V_4 in $\mathbf{E}$:

$$\mathbf{E} = [\mathrm{Embedding}([D_1;V_1;D_2;V_2;D_3;V_3;D_4]);\mathbf{H}^{'}] \tag{5}$$

Finally, we adopt the representation $\mathbf{E}$ as the input of multi-modal transformers.

4. Experiments

4.1. Datasets

We conduct experiments on 4 publicly benchmarks across 3 well-representative VrDU tasks, including document information extraction, document visual question answering, and document image classification. Specifically, the datasets involve FUNSD (Jaume et al., 2019), CORD (Park et al., 2019), RVL-CDIP (Rawat and Wang, 2017) and DocVQA (Mathew et al., 2021). The details and statistics (Table 1) of these datasets are as follows:

- **FUNSD**: FUNSD is a fundamentally document information recognition dataset which contains 4 entity types and 199 fully annotated forms, widely ranging from marketing, advertising and scientific reports. The forms are all one-page and rendered in a rasterized format with low resolution.

- **CORD**: CORD is a consolidated receipt dataset with box-level text and parsing class annotations. CORD consists of about 1000 receipts and the parsing word labels are divided into 30 different laebls.

- **RVl-CDIP**: RVL-CDIP is a large scale document image classification dataset sampled from IIT-CDIP (Lewis et al., 2006a) collection of tobacco litigation documents, containing 400,000 document images across 16 categories.

- **DocVQA**: We evaluate our approach on the DocVQA dataset for the document question answering. The dataset includes 12,767 document images of varied types and content, over 50,000 questions and answers.

Dataset	Field	Train	Vaild	Test
FUNSD	4	149	-	50
CORD	30	800	100	100
RVL-CDIP	16	320k	40k	40k
DocVQA	-	10,104	1,286	1,287

Table 1: Statistics of evaluation datasets.

4.2. Experiment Settings

We conduct our experiments in the few-shot settings. For each score in all experiments, we report the mean obtained with randomly seeds across the 5 different runs. For a fair comparison with previous work (Wang and Shang, 2022; **?**), we report the results of 2, 4, 6,and 8 shots for FUNSD and CORD and we then compute the standard micro precision, recall and F1 scores on both datasets. We evaluate our results under 1%, 5%, 10% and 20% of training samples for RVL-CDIP and DocVQA and report the accuracy and ANLS score respectively.

We adopt the state-of-art LayoutLMv3$_{LARGE}$ (Huang et al., 2022) as our backbone model and freeze the parameters of pre-trained model with only tuning the soft prompt embeddings and downstream classifiers. We also conduct and report the scores with LayoutLMv3$_{BASE}$ for a fair comparison with previous methods. We utilize the Spacy Entity Linker[1] to perform entity linking with one of the most popular and widely accessible knowledge bases, Wikidata[2]. Following previous work, the max length of total inputs is 512. We set the length of task soft prompts as 3 and knowledge soft prompts as 4. AdamW optimizer (Loshchilov and Hutter, 2017) is utilized to tune the parameters. For the prompt-tuning on the FUNSD dataset, we use a batch size of 2 and the learning rate of 1e-5. For the CORD dataset, we use a batch size of 2 and the learning rate of 1e-5. When tuning on the RVL-CDIP dataset, we set the batch size to 32 and the learning rate to 3e-5. Finally, on the DocVQA dataset, we use a batch size of 16 and the learning rate of 2e-5 for tuning our structural prompts and classifiers. We select the hyperparameters of each model via grid search on the validation sets. We choose the model with the best validation performance and report the scores on the test set.

4.3. Performance Evaluation

4.3.1. Document Information Extraction

The evaluation results on the FUNSD and CORD are shown in Table 2. We compare our methods with several state-of-the-art approaches:

- **RoBERTa** (Liu et al., 2019c) is a text-only pre-training language model which dynamically changes the masking pattern applied to the training samples with larger batches and more data.

- **LASER** (Wang and Shang, 2022) proposes a label-aware sequence-to-sequence framework with a novel labeling scheme to strengthen the label-region correlation.

- **LayoutLMv3** (Huang et al., 2022) firstly represent document images with linear projection features of image patches and proposes to facilitate cross-modal alignment learning through Word-Patch Alignment objectives. Without any special explanation, we adopt the LayoutLmv3$_{BASE}$ as the model backbones.

- **LAGER** (Krishnan et al., 2023) is based on LayoutLMv3 and utilizes the graph neural networks to capture the topological adjacency structures constructed from k-nearest bounding boxes to solve the image manipulations such as scaling, shifting or rotating the document images. LAGER introduce two heuristics to construct the node topology graph: k-nearest neighbors in space and k-nearest neighbors at multiple angles. Overall, the former heuristic is superior to latter and We report the scores of the model with the k-nearest neighbors in space rule.

From Table 2 we have several observations. First of all, our KnowVrDU$_{BASE}$ model achieves significant performance improvements compared with previous best results on both FUNSD and CORD datasets. It indicates that the model effectively incorporates the external task-specialized knowledge with our proposed unified structural prompts in document information extraction tasks. Moreover, it is worth mentioning that our KnowVrDU$_{LARGE}$ further outperforms the KnowVrDU$_{BASE}$ and other baseline methods, which we see the similar improvements on the precision and recall scores. It indicates that the more powerful pre-trained multi-modal transformer contains more prior general knowledge and factual descriptive information, which helps the model refine comprehension of our knowledge-aware prompts.

[1] https://github.com/egerber/spaCy-entity-linker
[2] https://www.wikidata.org/wiki/Wikidata

N	Method	FUNSD			CORD		
		Precision	Recall	F_1	Precision	Recall	F_1
2	RoBERTa	21.64±1.64	33.43±4.24	26.68±1.76	34.96±6.73	45.70±7.17	39.59±7.03
	LASER	30.40±4.89	35.20±7.20	32.36±5.14	-	-	-
	LayoutLMv3$_{BASE}$	44.29±6.14	58.96±7.20	50.43±6.03	47.21±6.25	58.99±4.94	52.41±5.85
	†LAGER$_{BASE}$	49.82±6.06	59.55±8.91	54.09±6.54	48.68±5.72	60.19±4.23	53.79±5.24
	KnowVrDU$_{BASE}$	50.37±5.85	59.61±6.26	54.42±6.13	49.36±5.96	59.84±3.28	54.29±4.75
	KnowVrDU$_{LARGE}$	**54.28±4.77**	**64.09±6.34**	**58.67±5.82**	**52.27±5.31**	**66.15±4.53**	**57.73±4.97**
4	RoBERTa	27.53±2.92	42.83±2.68	33.48±2.83	45.89±7.84	55.04±8.69	50.05±8.25
	LASER	44.91±2.42	50.25±3.26	47.36±2.18	-	-	-
	LayoutLMv3	65.32±3.89	77.97±2.26	71.06±3.04	54.18±5.01	64.92±3.76	59.04±4.53
	†LAGER$_{BASE}$	67.86±3.30	78.73±2.57	72.86±2.69	56.28±4.24	66.47±3.29	60.94±3.86
	KnowVrDU$_{BASE}$	69.20±3.03	78.51±2.64	73.67±2.31	56.69±4.13	68.16±4.05	62.24±5.01
	KnowVrDU$_{LARGE}$	**72.87±4.49**	**79.61±3.11**	**76.15±3.47**	**58.58±3.87**	**72.31±3.84**	**64.79±3.40**
6	RoBERTa	33.75±2.19	47.20±2.54	39.32±2.06	52.88±4.84	61.41±4.86	56.82±4.82
	LASER	48.64±2.14	53.54±2.10	50.96±1.95	-	-	-
	LayoutLMv3	71.19±3.75	80.83±1.09	75.68±2.58	60.91±3.51	69.16±2.76	64.76±3.16
	†LAGER$_{BASE}$	72.71±3.41	81.53±1.98	76.84±2.58	61.80±5.14	70.00±3.75	65.63±4.53
	KnowVrDU$_{BASE}$	72.23±3.26	83.25±1.16	77.31±2.36	63.27±2.76	72.90±4.68	67.59±3.38
	KnowVrDU$_{LARGE}$	**73.49±2.31**	**85.02±2.40**	**78.60±1.77**	**66.43±3.64**	**73.41±2.57**	**69.62±3.05**
8	RoBERTa	37.30±3.55	49.52±4.89	42.52±4.89	57.38±1.86	65.32±1.54	61.08±1.57
	LASER	-	-	-	-	-	-
	LayoutLMv3	74.31±2.19	81.75±2.60	77.85±2.29	64.49±3.24	72.21±2.17	68.12±2.77
	†LAGER$_{BASE}$	76.27±1.44	83.41±1.73	79.66±1.14	64.89±4.38	72.22±3.19	68.35±3.84
	KnowVrDU$_{BASE}$	78.14±1.68	84.65±1.29	81.44±1.42	66.17±2.83	72.99±3.71	69.23±3.16
	KnowVrDU$_{LARGE}$	**81.52±1.78**	**87.23±1.36**	**84.04±1.30**	**70.19±3.49**	**75.41±4.44**	**72.51±3.63**

Table 2: Performance of our KnowVrDU model and previous state-of-the-art models on the FUNSD and CORD datasets. N is the number of shots for FUNSD and CORD dataset. The best scores are in bold and the second-best scores are underlined. † marks scores produced by the model LAGER$_{nearst}$, which adopts the nearst heuristics to construct topology graph.

Ratio	**Model**	**RVL-CDIP**	**DocVQA**
1%	LayoutLMv3	29.70	16.84
	KnowVrDU$_{BASE}$	38.53	20.91
	KnowVrDU$_{LARGE}$	**45.31**	**25.42**
5%	LayoutLMv3	52.47	28.03
	KnowVrDU$_{BASE}$	56.90	31.78
	KnowVrDU$_{LARGE}$	**61.27**	**35.23**
10%	LayoutLMv3	71.14	42.11
	KnowVrDU$_{BASE}$	75.70	45.52
	KnowVrDU$_{LARGE}$	**78.22**	**49.67**
20%	LayoutLMv3	80.25	67.46
	KnowVrDU$_{BASE}$	84.03	69.81
	KnowVrDU$_{LARGE}$	**85.89**	**72.43**

Table 3: Results of KnowVrDU and the previous work on Document Image Classification and Document Question Answering tasks with different ratio settings.

4.3.2. Document Image Classification

We conduct our experiments on the RVL-CDIP dataset and report the average classification accuracy as the evalution metric. We adopt the start/end classifiers to predict the start/end position of document categories. A document classification is regarded as correct only if the predicted span is located in the labeled question span. We compare our method with LayoutLMv3 under 1, 5, 10 and 20 ratio settings and the evaluation results in shown in Table 3. From our preliminary experiments, we observe that our KnowVrDU$_{BASE}$ outperforms baselines on average improvements of 5.64%-15.61% in different sampling ratios. KnowVrDU$_{LARGE}$ achieves the new state-of-the-art performance, illustrating the effectiveness of our proposed structural prompts in document image classification.

4.3.3. Document Question Answering

We report the Average Normalized Levenshtein Similarity (ANLS) score which is widely used in question answering tasks, which greater ANLS scores reflect the stronger performance of models. The KnowVrDU$_{BASE}$ improves the ANLS scores compared with LayoutLMv3 and demonstrates the advantage of our framework in different ratio settings. The KnowVrDU$_{LARGE}$ further gains an absolute ANLS score of 4.97%-8.58% over

KnowVrDU$_{BASE}$ in distinct sampling ratios. The results show that KnowVrDU$_{BASE}$ is effective for the document visual question answering task.

4.4. Ablation Study

To further prove the effects of different components in KnowVrDU, we conduct the ablation study and report the experimental results as illustrated in Table 4. We conduct experiments from the 8-shot setting on the CORD dataset using the following ablation options:

- KnowVrDU$_{BASE}$ removes the **knowledge attention network**, which means that the noise in retrieved knowledge from open knowledge base are not reduced.
- KnowVrDU$_{BASE}$ removes the **knowledge integration module**, removing the whole "*Knowledge*" key-value pair in the structural prompt.
- KnowVrDU$_{BASE}$ removes the **soft prompts embedding** in the task indicator. Under this setting, we utilize the task name "*CORD*" to initialize the value of the task indicator.
- KnowVrDU$_{BASE}$ removes **all the components** proposed in our framework, which is essentially the same as native LayoutLMv3 model.

According to Table 4, we can see that the knowledge attention network constitutes a significant contribution to the model performance. It suggests that our framework can gain huge benefit from the refined task-specialized knowledge. Besides, by comparing the performance of KnowVrDU$_{BASE}$ w/o knowledge integration module, we claim that our knowledge-aware framework can reduce the gap between pre-trained model and downstream tasks through incorporating the external knowledge with knowledge retrieved modules. Moreover, the soft prompts also bring improvements to the model because they can enable the model to accommodate the unified prompt more effectively. Finally, the model without all three components produces the worst performance, which proves that our knowledge-aware prompt-tuning framework sufficiently enhances the task-specialized knowledge information and improves the performance of downstream tasks.

Method	Prec.	Rec.	F_1
KnowVrDU$_{BASE}$	**78.14**	**84.65**	**81.44**
w/o Attention	76.41	82.27	79.23
w/o Knowledge	74.92	81.66	78.14
w/o Soft Prompts	77.10	83.49	80.16
w/o All	74.31	81.75	77.85

Table 4: An ablation study of the knowVrDU$_{BASE}$ model on the CORD dataset.

4.5. Case Study

Figure 3 shows the comparison of the baseline and our KnowVrDU$_{BASE}$ model from the 8-shot setting on the FUNSD dataset. We observe that the baseline model fails to understand the semantics of the question and answer entities due to the limitation of the training set. Our method integrates the knowledge "*RECIPIENT: person or organization to whom a letter is addressed*" and "*FAX: telephone number of a facsimile line*" from the large-scale knowledge base. Therefore, KnowVrDU$_{BASE}$ recognizes the "*RECIPIENT*", "*FAX*" and recipient answer correctly. It indicates that our proposed knowledge-aware prompt-tuning framework could fuse the external task-specialized knowledge effectively and contribute to the visual document understanding properly.

5. Conclusion

In this paper, we propose a unified knowledge-aware framework for a wide applications of various downstream VrDU tasks. We propose to model heterogeneous VrDU structures through reformulating all tasks into extractive question answering tasks with task-specific prompts. We propose to reduce the knowledge gap through integrating external open-source knowledge to incorporate external open-source knowledge bases. Experimental results in few-shot settings demonstrate the effectiveness of our method on a wide variety of downstream VrDU tasks.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments on this paper. This work was supported by National Key R&D Program of China under Grant number 2021ZD0115801, the National Natural Science Foundation of China under Grant numbers 82090053 and the Tsinghua University Initiative Scientific Research Program of Precision Medicine under Grant number 2022ZLA0073. We would like to thank the Ant Group's support to the authors.

6. Bibliographical References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE inter-*

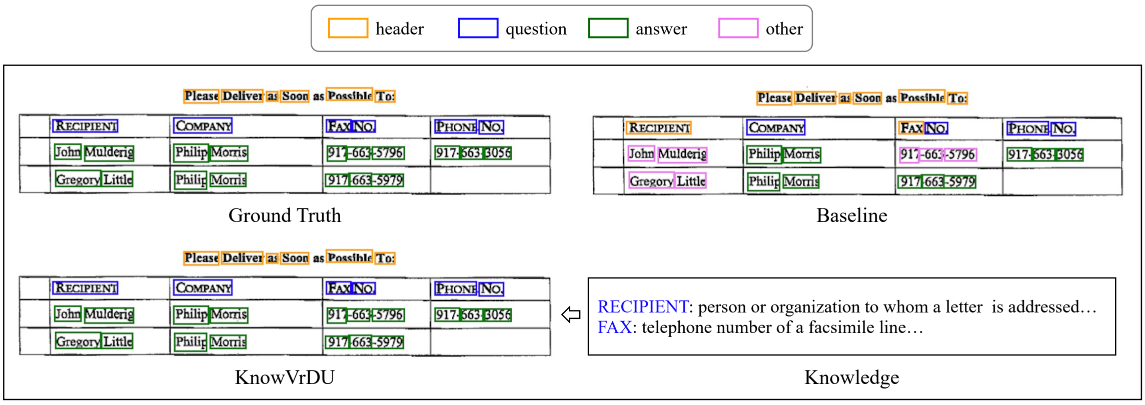


Figure 3: An example of the document from the FUNSD dataset and the predictions from the Baseline and KnowVrDU$_{BERT}$ model. We distinguish different entity types with different colors. The document header, question, answer and other type are represented by the orange, blue, green and violet color respectively. The Knowledge indicates the external specialized knowledge from open-source knowledge base.

national conference on computer vision, pages 2425–2433.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003.

Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Dar-Jen Chang, Ahmed H Desoky, Ming Ouyang, and Eric C Rouchka. 2009. Compute pairwise manhattan distance and pearson correlation coefficient of data points with gpu. In *2009 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing*, pages 501–506. IEEE.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document AI: benchmarks, models and applications. *CoRR*, abs/2111.08609.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4583–4592.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10767–10775.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4083–4091, New York, NY, USA. Association for Computing Machinery.

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

Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4459–4469.

Diederik P Kingma, Max Welling, et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Prashant Krishnan, Zilong Wang, Yangkun Wang, and Jingbo Shang. 2023. Towards few-shot entity recognition in document images: A graph neural network approach robust to image manipulation. *arXiv preprint arXiv:2305.14828*.

D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006a. 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.

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

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

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

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Andy T Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. 2022. Qaner: Prompting question answering models for few-shot named entity recognition. *arXiv preprint arXiv:2203.01543*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv:2103.10385*.

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

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019b. Graph convolution for multimodal information extraction from visually rich documents. *arXiv preprint arXiv:1903.11279*.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

J-L Meunier. 2005. Optimized xy-cut for determining a page reading order. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 347–351. IEEE.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

L. O'Gorman. 1993. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1162–1173.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, et al. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *arXiv preprint arXiv:2210.06155*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883.

Michael Shilman, Percy Liang, and Paul Viola. 2005. Learning nongenerative grammatical models for document analysis. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 962–969. IEEE.

Anikó Simon, J-C Pret, and A Peter Johnson. 1997. A fast algorithm for bottom-up document layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):273–277.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *stat*, 1050:20.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9154–9160.

J. Wang, L. Jin, and K. Ding. 2022. Lilt: A simple yet effective language-independent layout transformer for structured document understanding.

Zilong Wang and Jingbo Shang. 2022. Towards few-shot entity recognition in document images: A label-aware sequence-to-sequence framework. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4174–4186, Dublin, Ireland. Association for Computational Linguistics.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. 2022. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678.

Hao Wei, Micheal Baechler, Fouad Slimane, and Rolf Ingold. 2013. Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1220–1224.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

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

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

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.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Jun Zhao, Xin Zhao, WenYu Zhan, Tao Gui, Qi Zhang, Liang Qiao, Zhanzhan Cheng, and Shiliang Pu. 2022. Read extensively, focus smartly: A cross-document semantic enhancement method for visual documents NER. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2034–2043, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. 2022. Proqa: Structural prompt-based pre-training for unified question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4230–4243.

Shaowen Zhou, Bowen Yu, Aixin Sun, Cheng Long, Jingyang Li, Haiyang Yu, Jian Sun, and Yongbin Li. 2022. A survey on neural open information extraction: Current status and future directions. *arXiv preprint arXiv:2205.11725*.

7. Language Resource References

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe 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.

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.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

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*.

Waseem Rawat and Zenghui Wang. 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29:2352–2449.