Modeling Entities as Semantic Points for Visual Information Extraction in the Wild

Zhibo Yang\textsuperscript{1,2}\textsuperscript{*} Rujiao Long\textsuperscript{2,†} Pengfei Wang\textsuperscript{2}\textsuperscript{*} Sibo Song\textsuperscript{2} Humen Zhong\textsuperscript{2} Wenqing Cheng\textsuperscript{1} Xiang Bai\textsuperscript{1} Cong Yao\textsuperscript{2} \\
\textsuperscript{1}Huazhong University of Science and Technology \quad \textsuperscript{2}Alibaba Group

Abstract

Recently, Visual Information Extraction (VIE) has been becoming increasingly important in both the academia and industry, due to the wide range of real-world applications. Previously, numerous works have been proposed to tackle this problem. However, the benchmarks used to assess these methods are relatively plain, i.e., scenarios with real-world complexity are not fully represented in these benchmarks. As the first contribution of this work, we curate and release a new dataset for VIE, in which the document images are much more challenging in that they are taken from real applications, and difficulties such as blur, partial occlusion, and printing shift are quite common. All these factors may lead to failures in information extraction. Therefore, as the second contribution, we explore an alternative approach to precisely and robustly extract key information from document images under such tough conditions. Specifically, in contrast to previous methods, which usually either incorporate visual information into a multi-modal architecture or train text spotting and information extraction in an end-to-end fashion, we explicitly model entities as semantic points, i.e., center points of entities are enriched with semantic information describing the attributes and relationships of different entities, which could largely benefit entity labeling and linking. Extensive experiments on standard benchmarks in this field as well as the proposed dataset demonstrate that the proposed method can achieve significantly enhanced performance on entity labeling and linking, compared with previous state-of-the-art models. Dataset is available at \url{https://www.modelscope.cn/datasets/damo/SIBR/summary}.

1. Introduction

Visually Rich Documents (VRDs) are ubiquitous in daily, industrial, and commercial activities, such as receipts of shopping, reports of physical examination, product manuals, and bills of entry. Visual Information Extraction (VIE) aims to automatically extract key information from these VRDs, which can significantly facilitate subsequent processing and analysis. Due to its broad applications and grand technical challenges, VIE has recently attracted considerable attention from both the Computer Vision community [33, 34, 39] and the Natural Language Processing community [16, 35, 37]. Typical techniques for tackling this challenging problem include essential electronic conversion of image (OCR) [28, 30, 41], intermediate procedure of structure analysis [25] and high-level understanding of contents [35], among which entities play an important role as an aggregation of vision, structure, and language.

Though substantial progresses [11, 19, 35] have been made, it is still challenging to precisely and reliably extract key information from document images in unconstrained conditions. As shown in Fig. 1, in real-world scenarios documents may have various formats, be captured casually with
To highlight the challenges in real applications and promote the development of research in VIE, we establish a new dataset called Structurally-rich Invoices, Bills and Receipts in the Wild (SIBR for short), which contains 1,000 images with 71,227 annotated entity instances and 39,004 entity links. The challenges of SIBR lie in: (1) The documents are from different real-world scenarios, so their formats and structures might be complicated and varying; (2) The image quality may be very poor, i.e., blur, noise, and uneven illumination are frequently seen; (3) The printing process is imperfect that shift and rotation might happen.

To deal with these difficulties, we explore an novel approach for information extraction from VRDs. Different from previous methods, which usually employ a sequential pipeline that first uses an off-the-shelf OCR engine to detect and read textual information (location and content) and then fuses such information with visual cues for follow-up entity labeling (a.k.a. entity extraction) and linking in a multi-modal architecture (mostly a Transformer) [19, 37], the proposed method adopts a unified framework that all components, including text detection, text recognition, entity extraction and linking, are jointly modeled and trained in an integrated way. This means that in our method a separate OCR engine is no longer necessary. The benefits are two-fold: (1) The accuracy of entity labeling and linking will not be limited by the capacity of the OCR engine; (2) The running speed of the whole pipeline could be boosted.

Drawing inspirations from general object detection [4, 15, 40, 42] and vision-language joint learning [11, 14, 31], we put forward to model entities as semantic points (ESP for short). Specifically, as shown in Fig. 3, entities are represented using their center points, which are enriched with semantics, such as geometric and linguistic information, to perform entity labeling and linking. To better learn a joint vision-language representation, we also devise three training tasks that are well integrated into the paradigm. The entity-image text matching (EITM) task, which is only used in the pre-training stage, learns to align entity-level vision vectors and language vectors (encoded with off-the-shelf BERT) with a contrastive learning paradigm. Entity extraction (EE) and Entity linking (EL), the main tasks for VIE, are used in the pre-training, fine-tuning, and inference stages. In these two modules, region features and position embedding (from ground truth or detection branch) are encoded with transformer layers and then decoded to entity classes and relations. Owing to the joint vision-language representation, text recognition is no longer a necessary module in our framework, and we will discuss the impact of the text recognition branch in Sec. 5.4.

Extensive experiments have been conducted on standard benchmarks for VIE (such as FUNSD, XFUND, and CORD) as well as the proposed SIBR dataset. We found that compared with previous state-of-the-art methods, the proposed ESP algorithm can achieve highly competitive performance. Especially, it shows an advantage in the task of entity linking. Our main contributions can be summarized as follows: (1) We curate and release a new dataset for VIE, in which the document images are with real-world complexity and difficulties. (2) We devise a unified framework for spotting, labeling and linking entities, where a separate OCR engine is unnecessary. (3) We adopt three vision-language joint modeling tasks for learning informative representation for VIE. (4) Extensive experiments demonstrate the effectiveness and advantage of our approach.

2. Related Work

The proposed approach has been inspired by ideas from object detection and document understanding. Exemplar approaches include detection with grouping [24], text spotting [23, 43] and information extraction joint training [34], and vision-language pre-training for VIE [35].

Detection with Grouping Methods consider an object as a collection of parts arranged in specific relations. Object as points methods [4, 15, 40] firstly detected four corner points and one center point of an object and then grouped them with the geometric-based algorithm or associative embedding features. SegLink [27] and its variant SegLink++ [32] took a small segment of a text instance as the fundamental element and linked them together to form the bounding boxes. In [8, 9], hierarchical part-whole ideas were approached at the model level and feature space. Visual entity labeling and linking can also be learned in this manner to some extent. We model visual entities as points and group them with their semantic relations.

Vision-Language Pre-training Methods usually extract visual document information by using a 2D positional encoding of BERT-like architecture. The innovation of these works lies in novel pre-training tasks or new multi-modal transformer layers. LayoutLM [35] and its variant LayoutLM-v2 [37] and layoutLM-v3 proposed several novel pretraining tasks, including Masked Visual-Language Model, Text-Image Alignment, Text-Image Matching, and Word-Patch Alignment. StrucText [19] introduced a multi-granularity image and text alignment for entity labeling. BROS [10] proposed an area-masking self-supervision strategy that reflected the 2D nature of text blocks. In addition to pre-training tasks, DocFormer [1] and SelfDoc [18] also introduced cross-attention layers for better fusing of visual and language features. LiLT [33] extended document understanding to multi-languages by designing language-independent attention layers. The above methods usually depended on offline OCR engines, which may lead to error amplification and redundant computation.

Jointly Training Methods aim at improving information
Figure 2. Comparisons of color distribution across different datasets. A distribution of background from 5 to 250 is for better readability. (Best viewed in color.)

Table 1. Comparison between SIBR and other datasets. “Scan” is short for scanned receipts or documents, “CamC” is short for Camera-captured images. “Overlap” showcases the proportion of images with overlapping entity boxes.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Box</th>
<th>Link</th>
<th>Language</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>SROIE</td>
<td>45279</td>
<td>9529</td>
<td>97012</td>
<td>95%</td>
</tr>
<tr>
<td>FUNSD</td>
<td>10420</td>
<td>102291</td>
<td>7470</td>
<td>20%</td>
</tr>
<tr>
<td>SIBR</td>
<td>1139</td>
<td>199</td>
<td>1393</td>
<td>50%</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>30%</td>
</tr>
</tbody>
</table>

4. Methodology

4.1. Overall Architecture

The goal of a VIE algorithm is to learn to locate key text contents and extract entity labels and linking information from visually-rich documents. We approach VIE from a point-based perspective, i.e., we learn to represent each entity, its label and relations with other entities by a single point. Specifically, we adopt the CenterNet framework [40], owing to its simple and concise pipeline. By utilizing the unified paradigm for feature learning, the model can benefit from joint learning across different modalities and tasks.

Visual Backbone We choose ConvNeXt-tiny-FPN as the visual backbone for its excellent trade-off between performance and efficiency. ConvNeXt [22] is a pure CNN architecture, which compares favorably with state-of-the-art hierarchical vision Transformers across multiple vision
benchmarks while retaining the simplicity and efficiency of ConvNets. Moreover, we adopt feature pyramid network (FPN) [20] to fully exploit multi-scale visual features. Let \( I \in \mathbb{R}^{w \times h \times 3} \) be the input image of width \( w \) and height \( h \). The visual backbone then produces a feature map \( F_2 \in \mathbb{R}^{\frac{w}{4} \times \frac{h}{4} \times 256} \) as the input to the EAP module.

**Entities-As-Points (EAP)** Given an image, CenterNet [40] extracts a set of bounding boxes for each object class with the help of a heatmap head, a size head and an offset head. We build the EAP module upon the original CenterNet framework with the following architectural modifications.

First, we devise the size head into a quadrangle head, which predicts the location of four vertices to detect arbitrary-oriented entity text instances. Therefore, the total text detection loss is computed as follows:

\[
L_{\text{Det}} = L_{\text{Heatmap}} + L_{\text{Quad}} + L_{\text{Offset}}
\]

Second, we propose three feature transformation branches to encode cross-modal features, entity type features, and entity relational features respectively. Each feature transformation branch consists of a \( 3 \times 3 \) convolutional layer, a ReLU layer, and a final \( 1 \times 1 \) convolutional layer. Moreover, the transformed feature maps are extracted based on the predicted points from the text detection branch. In this way, cross-modal features and entity-related features are produced in a unified point-based representation. Third, we adopt a simplified version of CRNN [29] as the text recognition branch. The predicted entity boxes will be fed into a grid-sample module to sample the text features \( F_1 \in \mathbb{R}^{N \times C \times S_h \times S_w} \). \( S_h \) and \( S_w \) denote the sampling size of height and width respectively while \( N \) denotes the number of text segments. Our CRNN variant is composed of 6 convolutional layers for feature extraction, 1 bi-directional LSTM layer for sequential modeling, and 1 linear projection layer for label prediction. Connectionist Temporal Classification (CTC) loss \([5]\) is adopted as recognition loss \( L_{\text{rec}} \) to align sequence predictions with the labels. In fact, our model does not depend on the recognition module, which is only used to output the end-to-end results, so we did not use the recognition results on FUNSD, XFUND, and CORD.

Overall, the point-based nature of entity representation possesses several advantages. First, our proposed method learns visual, textual, and layout representations simultaneously on a grid-based feature map, avoiding sophisticated serialization steps as for sequence-based methods \([11,35,36]\), especially in real-world scenarios (e.g., misaligned second-time printing, receipts with complex layout) where such linearization step is nontrivial. Second, the text is usually truncated with a preset maximum number of tokens for sequence-based approaches. Nevertheless, our proposed model is not limited by a token number and can learn more holistic representations to mine inter-text relationships without any text truncation.

**Entity-Image Text Matching (EITM)** Recent progress in VIE has been mainly powered by multi-modal pre-training. Therefore, we harness an entity-image text matching module during the pre-training stage to enhance cross-modal feature learning. An image-text matching task is adopted to encourage cross-modal interactions and mutual alignment between visual and textual representations. EITM is only applied to the pre-training stage, thus enabling recognition-free prediction for entity labeling and linking. We employ the idea of contrastive learning to mutually align entity image and entity text representations into semantic space. It aims to find the closest entity image embedding from a batch given an entity text embedding. Similarly, for a given entity image embedding, the objective is to find the closest entity text embedding. This results in an InfoNCE loss over each batch of entity image-text pairs for pre-training our model.
\[ L_{v2t} = - \sum_{j=1}^{N} \log \frac{\exp(v_i \cdot t_j \div \tau)}{\sum_{k=1}^{N} \exp(v_i \cdot t_k \div \tau)} \]  

where \( N \) denotes the number of image-text pairs which equals the number of text segments, \( v_i \) and \( t_j \) represent \( i \)-th entity image embedding and entity text embedding respectively, and \( \tau \) denotes the hyperparameter of temperature. Similarly, for each \( t_i \), the loss is formulated as,

\[ L_{t2v} = - \sum_{j=1}^{N} \log \frac{\exp(t_i \cdot v_j \div \tau)}{\sum_{k=1}^{N} \exp(t_i \cdot v_k \div \tau)} \]  

Finally, the total loss for EITM is defined as,

\[ L_{EITM} = L_{v2t} + L_{t2v} \]  

**Entity Extraction (EE)** To strengthen the entity type learning, we first concatenate vision-language features \( F_1 \) and entity type features \( F_2 \) to leverage both visual and semantic representations.

We then employ two self-attention layers to aggregate features from all entities. Last, the aggregated features are fed into the entity classifier, which is a ReLU-fc-ReLU-fc-Sigmoid block. We set the hidden size, number of self-attention heads, feed-forward/filter size, and number of Sigmoid block. We set the hidden size, number of self-attention heads, feed-forward/filter size, and number of Sigmoid block.

Following [19], we use both DocBank and RVL-CDIP for pre-training. DocBank contains 500K images, of which

\[ L_{EE} = \begin{cases} 
-(1 - p_i)\gamma \log(p_i), & p_i^* = 1 \\
-p_i^* \log(1 - p_i), & \text{otherwise} 
\end{cases} \]  

where \( p_i \) is the predicted probability vector and \( p_i^* \) is the ground truth. We set \( \gamma \) to 4.

**Entity Linking (EL)** In many VIE applications, an entity often contains multiple text segments or text boxes. Therefore, it is important to model intra-entity links in addition to inter-entity links. Inter-entity links rely more on the spatial correlations of text segments, however, intra-entity links rely heavily on text semantics. Thus, we concatenate the cross-modal feature \( F_1 \) with entity relational feature \( F_3 \) to handle the lack of semantics and apply two layers’ self-attention modules to capture inter-text dependencies.

The structure of the module is the same as Entity Extraction Module. The only difference is that the Entity Linking Module constructs a relationship matrix between entity text instances. We denote the output of the self-attention layers by \( E_i' \in \mathbb{R}^{N_i \times C} \), where \( C \) is the dimension of input feature embeddings and \( N_i \) is the number of key-value pairs. First, we broadcast \( E_i' \) to obtain \( E_{i1} \in \mathbb{R}^{1 \times N_i \times d} \) and \( E_{i2} \in \mathbb{R}^{N_i \times 1 \times d} \). Since entity links are vectors with directions, we formulate the relational matrix via element-wise subtraction as given in Eq. (6) to encourage learning directed link representations,

\[ E_i^* = E_{i1} - E_{i2}, \quad E_i^* \in \mathbb{R}^{N_i \times N_i \times C} \]  

\( E_i^* \) is input into the classifier to obtain link results. The classifier is a ReLU-fc-ReLU-fc-Sigmoid block. Since the entity linking matrices are usually sparse, we use Focal loss as the same as the entity labeling task and denote as \( L_{EL} \).

**4.2. Training Objectives**

**Pre-Training** The pre-training is conducted on DocBank and RVL-CDIP datasets with all aforementioned modules. Unlike previous methods that focus on designing mask-based or cross-alignment tasks, we generate pseudo-labels to mine semantic and relational information in these pre-training datasets. See Sec. 5.1 for more details. Thus, the overall pre-training objective is,

\[ L_{PT} = L_{Det} + L_{EITM} + \lambda_{EL} L_{EL} + \lambda_{EE} L_{EE} \]  

**Fine-Tuning** Since EITM task only exists in the pre-training stage and recognition tasks only exist in fine-tuning, the loss term for fine-tuning is,

\[ L_{FT} = L_{Det} + L_{Det}^* + \lambda_{EL} L_{EL} + \lambda_{EE} L_{EE} \]  

where \( \lambda_{EL} \) is set to 1 for pre-training and 10 for fine-tuning and \( \lambda_{EE} \) is set to 10 for pre-training and 200 for fine-tuning. It is worth noting that, the recognition branch in our framework is optional, i.e., it can be removed or replaced by any off-the-shelf text recognizer given entity images.

**5. Experiments**

**5.1. Datasets**

DocBank and RVL-CDIP are used for pre-training, while the FUNSD [13], XFUND [34] and CORD [26] are adopted to compare with SOTA methods to verify the effectiveness of the ESP on EE and EL tasks.

Following [19], we use both DocBank and RVL-CDIP for fine-tuning. DocBank contains 500K images, of which
400K are for training, 50K for validation, and 50K for testing. RVL-CDIP includes 400K grayscale images in 16 categories, with 25K images per category. 320K are for training, 40K for validation, and 40K for testing.

**FUNSD** is composed of 149 training samples and 50 testing samples, where the images are scanned English documents. **XFUND** extends the FUNSD dataset to 7 other languages, with the same number of images in each language as the FUNSD. **CORD** consists of 1000 fully annotated receipts, including 800 training samples, 100 validation samples, and 100 testing samples. The FUNSD, XFUND, and CORD are annotated with both word and block locations. Combining coordinate information of text words and blocks, the segment-level annotation is generated for fine-tuning the ESP model.

**Pre-Training Pseudo Label Generation** The existing pre-training methods mainly focus on designing self-supervised tasks to boost performance, such as masked language models or word-patch alignment. We approach this by exploiting knowledge from pre-training datasets. The pipeline to generate pseudo-labels for pre-training EE and EL tasks of key-value pairs is described as follows:

First, the key library is generated by counting the occurrences of text contents. The number of occurrences more than 100 and 15 will be considered as key in DocBank and RVL-CDIP, respectively. Generally, the keys should not be too long or too short, thus the text, with more than 50 characters or less than 3, is skipped. Second, the key-value attributes are actually the pseudo labels of the EE task, and the key-value relationships are the pseudo labels of the EL task.

For the key-value pairs, the key/value attributes are actually the pseudo labels of the EE task, and the key-value relationships are the pseudo labels of the EL task.

The visualization of pseudo-label is shown in Fig. 4. For there are some errors, the EL task is only trained with matched key/value links and the others are masked in the pre-training.

### 5.2. Implementation Details

**Pre-Training** We pre-train the ESP models for 10 epochs with an initial learning rate of 1.25e-4, which is decayed by a factor of 0.1 at the 5th and 8th epoch, respectively. The longer side of images is resized to 1024, with the aspect ratio kept. The OCR information, obtained from PaddleOCR [17], is adopted to generate pseudo labels, including key-value labeling and linking information with the algorithm mentioned in Sec. 5.1.

**Fine-Tuning** Keeping the input size the same as the pre-training stage, the batch size is set to 8 and the initial learning rate is 1.25e-4. Since there are only 149 training samples in FUNSD, we train our model for 300 epochs, and the learning rate drops to 1.25e-5 and 1.25e-6 at the 200th and 250th epochs, respectively. CORD and XFUND are trained for 150 epochs, and the learning rate drops to 1.25e-5 and 1.25e-6 at the 100th and 130th epochs.

**Evaluation** Following [19], we aggregate segment features of the text into one entity if the entity has multiple segments on FUNSD, XFUND, and CORD. But when testing end-to-end on SIBR, model inference does not rely on any annotation information. After decoding, the boxes, categories, and links are directly matched with GT. We use F1-score as our evaluation metrics for both EE and EL tasks.

---

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>Pretrain Data Size</th>
<th>FUNSD EE</th>
<th>FUNSD EL</th>
<th>XFUND EE</th>
<th>XFUND EL</th>
<th>CORD EE</th>
<th>CORD EL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT_BASE [35]</td>
<td>110M + α</td>
<td>-</td>
<td>60.92</td>
<td>27.65</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>89.68</td>
</tr>
<tr>
<td>RoBERTa_BASE [2]</td>
<td>125M + α</td>
<td>-</td>
<td>66.48</td>
<td>-</td>
<td>-</td>
<td>47.69</td>
<td>93.54</td>
<td></td>
</tr>
<tr>
<td>LayoutLM_BASE [35]</td>
<td>160M + α</td>
<td>11M</td>
<td>88.41</td>
<td>45.86</td>
<td>-</td>
<td>96.07</td>
<td>95.21</td>
<td></td>
</tr>
<tr>
<td>SelfDoc [18]</td>
<td>-</td>
<td>0.32M</td>
<td>83.36</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UDoc [6]</td>
<td>272M + α</td>
<td>11M</td>
<td>87.93</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>StrucText [19]</td>
<td>107M + α</td>
<td>0.9M</td>
<td>83.09</td>
<td>44.10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LayoutLMv2_BASE [37]</td>
<td>200M + α</td>
<td>11M</td>
<td>82.76</td>
<td>42.91</td>
<td>-</td>
<td>94.95</td>
<td>95.59</td>
<td></td>
</tr>
<tr>
<td>LiLT_BASE [33]</td>
<td>-</td>
<td>11M</td>
<td>85.74</td>
<td>62.76</td>
<td>85.85</td>
<td>81.25</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>XLYayout_BASE [7]</td>
<td>345M + α</td>
<td>30M</td>
<td>83.35</td>
<td>-</td>
<td>82.04</td>
<td>67.79</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BROS_BASE [10]</td>
<td>110M + α</td>
<td>11M</td>
<td>83.05</td>
<td>71.46</td>
<td>-</td>
<td>96.50</td>
<td>95.73</td>
<td></td>
</tr>
<tr>
<td>LayoutLMv3_BASE [11]</td>
<td>133M + α</td>
<td>11M</td>
<td>90.29</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>96.56</td>
<td></td>
</tr>
<tr>
<td>BERT_LARGE [35]</td>
<td>340M + α</td>
<td>-</td>
<td>65.63</td>
<td>29.11</td>
<td>-</td>
<td>90.25</td>
<td>94.31</td>
<td></td>
</tr>
<tr>
<td>RoBERTa_LARGE [2]</td>
<td>355M + α</td>
<td>-</td>
<td>70.72</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>93.80</td>
<td></td>
</tr>
<tr>
<td>LayoutLM_LARGE [35]</td>
<td>343M + α</td>
<td>11M</td>
<td>77.89</td>
<td>42.83</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>95.41</td>
</tr>
<tr>
<td>LayoutLMv2_LARGE [37]</td>
<td>426M + α</td>
<td>11M</td>
<td>84.20</td>
<td>70.57</td>
<td>-</td>
<td>-</td>
<td>96.01</td>
<td>97.29</td>
</tr>
<tr>
<td>BROS_LARGE [10]</td>
<td>340M + α</td>
<td>11M</td>
<td>84.52</td>
<td>77.01</td>
<td>-</td>
<td>-</td>
<td>97.28</td>
<td>97.40</td>
</tr>
<tr>
<td>DocFormer_LARGE [1]</td>
<td>536M + α</td>
<td>5M</td>
<td>84.55</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>96.99</td>
<td></td>
</tr>
<tr>
<td>LayoutLMv3_LARGE [11]</td>
<td>368M + α</td>
<td>11M</td>
<td>92.08</td>
<td>80.35</td>
<td>-</td>
<td>-</td>
<td>97.46</td>
<td>98.28</td>
</tr>
<tr>
<td>ESP (ours)</td>
<td>50M</td>
<td>0.9M</td>
<td>91.12</td>
<td>88.88</td>
<td>89.13</td>
<td>92.31</td>
<td>95.65</td>
<td>98.80</td>
</tr>
</tbody>
</table>

Table 2. Comparisons on Entity Extraction and Entity Linking tasks. α: parameters for OCR engine should be considered.
Ground Truth (C) ESP (B) LayoutXLM

Figure 5. Qualitative results of ESP on the proposed dataset. The columns from left to right represent the visualization of ground truth, LayoutXLM, and ESP. The rectangles in green and red stand for the results of correct and incorrect information extraction, respectively.

The LayoutXLM is more error-prone, due to the lower quality of OCR results. Only key-value pairs are shown for better visualization.

Table 3. Evaluation on FUNSD and XFUND (fine-tuning on X, testing on X). Note that, LayoutXLM adopts language-specific fine-tuning and LiLT applies a plain text model (InfoXLM). However, ESP does not involve any multilingual language models as EITM is frozen during the fine-tuning stage.

5.3. Comparison with the State-of-the-Arts

In this section, we compare ESP with LayoutLM-based models [11, 35–37], BERT [3], RoBERTa [2], XYLay- out [7], StrucText [19], and BROS [10] on three public benchmarks, and the results are summarized in Tab. 2.

We observe that ESP outperforms all existing methods on the EL task, and it boosts the upper bound of FUNSD and XFNSD to new record (88.88% and 92.31%). This proves that visual and geometric features can largely benefit entity linking prediction. On the EE task, ESP also achieves competitive performance, but its performance is slightly lower than that of LayoutLMv3 on FUNSD and CORD. Note that during evaluation LayoutLMv3 additionally used the ground truth of text contents, while ESP did not, so the comparison is not totally fair. Moreover, the model size (50M parameters) and the image set for pre-training (0.9M samples) of ESP are much smaller, compared with LayoutLMv3 (over 368M parameters and 11M samples).

We also conduct experiments on XFUND and FUNSD to verify the importance of visual features, and the results are shown in Tab. 3. Previous LayoutLM and its variants extract image feature by additional un-trainable visual encoders, leading to inevitably losses of OCR related details. We argue that fine-grained visual cues, such as fonts, colors, positions and typographic information, are crucial for VIE.

We compare ESP with three multi-language approaches, LayoutXLM, RoBERTa, and LiLT. We pre-train ESP on English datasets of DocBank and RVL-CDIP, but fine-tune and test it on language-specifically FUNSD (English) and XFUND (7 other languages). Note that the EITM task is not trained in the fine-tuning stage, meaning that the model does not perceive the other 7 kinds of languages. We find that ESP maintains at a plateau of 87.3% on XUFNSD, much better than language-independent LiLT, which demonstrates the effectiveness of visual information extractors.

5.4. Ablation Study

Effect of Pre-training Tasks This group of experiments is conducted on FUNSD to illustrate the impact of various...
Table 5. Evaluations and ablation studies on SIBR. “Det” and “Rec” denote “detection” and “recognition”, respectively. The recognition of text is considered correct when the normalized edit distance is less than 0.3. The † denotes a co-training process with text recognition branch.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Task</th>
<th>Model</th>
<th>EE</th>
<th>EL</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>EE/EL</td>
<td>TRIE [39]</td>
<td>85.62</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LayoutXLM [36]</td>
<td>94.72</td>
<td>83.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ESP¹</td>
<td>94.93</td>
<td>85.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ESP</td>
<td>95.27</td>
<td>85.96</td>
</tr>
<tr>
<td>(b)</td>
<td>Det + EE/EL</td>
<td>TRIE [39]</td>
<td>46.40</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LayoutXLM [36]</td>
<td>75.07</td>
<td>53.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ESP¹</td>
<td>91.69</td>
<td>73.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ESP</td>
<td>92.29</td>
<td>74.44</td>
</tr>
<tr>
<td>(c)</td>
<td>Det + Rec + EE/EL</td>
<td>TRIE [39]</td>
<td>68.55</td>
<td>46.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LayoutXLM [36]</td>
<td>67.76</td>
<td>48.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ESP¹</td>
<td>70.45</td>
<td>51.47</td>
</tr>
</tbody>
</table>

We have released an in-the-wild document dataset for VIE, in which the images are challenging in their diverse appearances and complex structures. To deal with these difficulties, we further propose a unified framework to model entities as semantic points, in which entity spotting, labeling, and linking can be learned in an end-to-end manner. As for LayoutXLM, it is trained with OCR-matched truth and evaluated with ground truth. In order to conduct experiments on the third condition (c), we use Duguang OCR engine¹ to get character-level locations and contents which are not provided by PaddleOCR. We closely follow the experimental setup of TRIE² and LayoutXLM². The ablation studies under different restrictions are shown in Tab. 5.

(a) With GT Input. ESP achieves SOTA results on both EE and EL tasks, outperforming the other two methods by 0.55% and 1.97%, respectively. This proves that ESP can effectively handle the challenges in SIBR.

(b) With Predicted Boxes. In the second comparison, the performance of ESP exceeds the other two methods by a large margin. Specifically, taking the detection errors of PaddleOCR into consideration, the performance of LayoutXLM decreases by 19.65% in EE and 30.36% in EL. The performance of TRIE decreases by a larger margin. However, ESP only decreases by 2.98% and 11.52%, respectively, demonstrating its robustness to detection error.

(c) With Predicted Boxes and Text. Taking the recognition results into consideration, two of the three methods’ F1-scores decrease by a significant margin (>20%). Since TRIE gives an egregious result, we leave the corresponding result blank. We have several conclusions in this experiment: (1) The advantages of ESP become smaller with predicted text contents, (2) The text recognition branch is not easily trained with small data, (3) The connection between OCR and VIE should be studied more. The ultimate goal of VIE is to get not only the correct entity labels but also the correct contents. A qualitative comparison is shown in Fig. 5. Compared with LayoutXLM, ESP can better handle challenges such as overlap and printing shift.

Effect of Text Recognition We conduct the following experiments to discuss the effectiveness of the optional text recognition branch, and the results are presented in Tab. 5. Specifically, taking the recognition branch of ESP with a text recognition branch. It can be observed from the first comparison that the result of entities slightly decreases by 0.34% with a recognition branch. The reason for the drop is the insufficient fine-tuning of text recognition on small data. Besides, we presume that the pre-training tasks have learned a better vision-language representation, which means explicit text information is not essential to the proposed ESP.

6. Conclusion and Future Work

We have released an in-the-wild document dataset for VIE, in which the images are challenging in their diverse appearances and complex structures. To deal with these difficulties, we further propose a unified framework to model entities as semantic points, in which entity spotting, labeling, and linking can be learned in an end-to-end manner. We also introduce three pre-training tasks to enforce the network to learn a better vision-language representation. Extensive experiments demonstrate that the proposed method can achieve significantly enhanced performance. The proposed ESP can also be extended to more general VIE problems, which we will leave for future research.

Acknowledgement. This work was supported by the National Natural Science Foundation of China 62225603.
References


