MatchVIE: Exploiting Match Relevancy between Entities for Visual Information Extraction

Guozhi Tang[1]*, Lele Xie[3]*, Lianwen Jin[1,2]†, Jiapeng Wang[1], Jingdong Chen[3], Zhen Xu[4], Qianying Wang[4], Yaqiang Wu[4] and Hui Li[4]
[1]School of Electronic and Information Engineering, South China University of Technology, China
[2]Guangdong Artificial Intelligence and Digital Economy Laboratory (Pazhou Lab), Guangzhou, China
[3]Ant Group, China
[4]Lenovo Research, China
{eetanggz, eejpwang}@mail.scut.edu.cn, eelwjin@scut.edu.cn,
{yule.xll, jingdongchen.cjd}@antgroup.com, {xuzhen8,wangqya,wuyqe,lihuid} @lenovo.com

Abstract

Visual Information Extraction (VIE) task aims to extract key information from multifarious document images (e.g., invoices and purchase receipts). Most previous methods treat the VIE task simply as a sequence labeling problem or classification problem, which requires models to carefully identify each kind of semantics by introducing multimodal features, such as font, color, layout. But simply introducing multimodal features couldn't work well when faced with numeric semantic categories or some ambiguous texts. To address this issue, in this paper we propose a novel key-value matching model based on a graph neural network for VIE (MatchVIE). Through key-value matching based on relevancy evaluation, the proposed MatchVIE can bypass the recognitions to various semantics, and simply focuses on the strong relevancy between entities. Besides, we introduce a simple but effective operation, Num2Vec, to tackle the instability of encoded values, which helps model converge more smoothly. Comprehensive experiments demonstrate that the proposed MatchVIE can significantly outperform previous methods. Notably, to the best of our knowledge, MatchVIE may be the first attempt to tackle the VIE task by modeling the relevancy between keys and values and it is a good complement to the existing methods.

1 Introduction

The Visual Information Extraction (VIE) aims to extract key information from document images (invoices, purchase receipts, ID cards, and so on), instead of plain texts. The particularity of the VIE task brings several additional difficulties. Firstly, documents usually have diverse layouts, which vary significantly even for the same type of the document (e.g., the invoices from different vendors). In addition, the documents may contain multiple similar but not identical texts (e.g., the issue date and the expiration date) that are very difficult to distinguish.

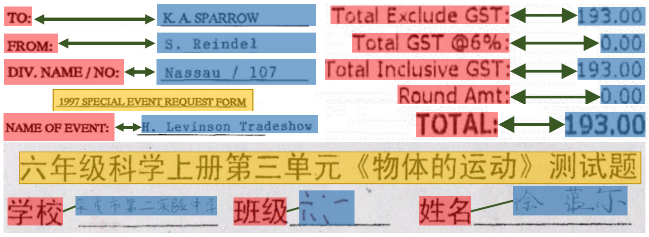

TO:
K. A. SPARROW
FROM:
S. Reindel
DIV. NAME / NO:
Nassau / 107
1997 SPECIAL EVENT REQUEST FORM
NAME OF EVENT:
H. Levinson Tradeshow
Total Exclude GST:
193.00
Total GST @6%:
0.00
Total Inclusive GST:
193.00
Round Amt:
0.00
TOTAL:
193.00
六年级科学上册第三单元《物体的运动》测试题
学校
班级
六、一
姓名

Figure 1: The entity categories of the blue text segments (value) can be identified according to the semantics of connected red text segments (key). The entity categories of the yellow text segments can be identified according themselves semantics.

How to leverage effectively both semantic features and visual features has become the focus of recent studies. Some methods have attempted to incorporate the positions of the texts [Jiang *et al.*, 2019b; Hwang *et al.*, 2019], while others [Qian *et al.*, 2019; Liu *et al.*, 2019] have modeled the layout information through Graph Neural Network (GNN) [Kipf and Welling, 2016; Veličković *et al.*, 2017]. Some methods [Xu *et al.*, 2020; Yu *et al.*, 2020; Zhang *et al.*, 2020] utilized the Convolutional Neural Network (CNN) [He *et al.*, 2016; Huang *et al.*, 2017] to fetch image features and fused them with semantic features for improving performance. These methods have achieved improved results by considering the visual features. However, most previous methods are confined to sequence labeling or direct classification, which requires models to assign each entity with corresponding labels carefully when facing numerous semantic categories or some ambiguous texts. In visual-rich documents, the layout information between entities is an important reasoning clue. As shown in Figure 1, the entity categories of **blue text segments (value)** can be identified according to the semantic of **red text segments (key)**.

Compared with sequence labeling or direct classification, we find that studying the relevancy between keys and values could be an another effective solution for the VIE, based

[*] These authors contributed equally to this work.
[†] Corresponding author: Lianwen Jin.

on the following observations and considerations: **(1)** The texts in document images usually appear in the form of key-value pairs. If the corresponding key can be found for a specific value, the attribute (category) of this value can be naturally determined. **(2)** There may be multiple similar texts in one document image (e.g., registration, amount and expiry dates), and keys of these values can help the model distinguish among them. **(3)** Considering the relevancy between keys and values could significantly simplify the learning process for models and bypass the recognition to similar semantic. **(4)** As for the standalone texts (without keys), they are easy to identify by semantics. This is also why these values can appear independently in the document images.

Therefore, in this paper we propose a novel key-value matching model for VIE (MatchVIE). It can effectively integrate the semantic, location and visual information of entities and innovatively consider the edge relationship of the graph network to evaluate the relevance of entities. There are two branches in the network, one is to measure the relevancy between keys and values, and the other auxiliary branch is to process the standalone texts by combining sequence labeling. To the best of our knowledge, the proposed MatchVIE may be the first attempt to accomplish the VIE task by modeling the relevancy between keys and values. The main contributions of our study are summarized as follows:

- Our MatchVIE effectively incorporates features (including layout, entity relevance, etc.) and brings significant accuracy improvements and surpasses existing methods.
- Our MatchVIE bypasses the recognition to various similar semantics by focusing only on the strong relevancy of entities. This simplifies the process of VIE.
- We introduce a simple yet effective operation named Num2Vec to tackle the instability of encoded values, which helps model converge more smoothly.
- Most importantly, this paper demonstrates that the visual information can be effectively extracted through modeling the relevancy between keys and values, which provides a brand-new perspective to solve the VIE task.

2 Related Works

Recently, VIE has attracted the attention of many researchers. Researchers are of the view that image features of documents are very useful because features are mixed representations of fonts, glyph, colors, etc. As the VIE task involves the document images, some researchers regard it as a pure computer vision task, such as EATEN[Guo *et al.*, 2019], TreyNet[Carbonell *et al.*, 2020] and [Boroş *et al.*, 2020]. These methods solved the VIE task from the perspective of Optical Character Recognition (OCR). For each type of entities, these methods designed corresponding decoders that were responsible for recognizing the textual content and determining its category. This method couldn't work well when faced with complex layouts because of the absence of semantic features.

The core of the research is how to make full use of the multimodal features of document images. Researchers have approached VIE from various perspectives. [Hwang *et al.*, 2019] and [Jiang *et al.*, 2019b] serialized the text segments based on the coordinate information and fed coordinates to a sequence tagger. However, simply treating the position as some kind of features might not fully exploit the visual relationships among texts. To make full use of the semantic features and position information, Chargrid [Katti *et al.*, 2018] mapped characters to one-hot vectors which filled the character regions on document images. An image with semantic information is fed into a CNN for detection and semantic segmentation to extract entities. The later BERTgrid [Denk and Reisswig, 2019] followed a similar approach but utilized different word embedding methods. However, it introduced a vast amount of calculation by using channel features to represent semantics, especially languages with large categories.

Therefore, it is usually a better solution to construct a global document graph, using semantic features as node features and the spatial location features of text segments as edge features. Several methods [Qian *et al.*, 2019; Liu *et al.*, 2019; Yu *et al.*, 2020; Gal *et al.*, 2020; Cheng *et al.*, 2020] employed the GNN to model the layout information of documents. Through the messages passing between nodes, these models could learn the overall layout and distribution of each text, which was conductive to the subsequent entity extraction. For example, [Gui *et al.*, 2019] proposed a lexicon-based graph neural network that treated the Chinese NER (Named Entity Recognition) as a node classification task. Besides, the GraphIE [Qian *et al.*, 2019] and the model proposed by [Liu *et al.*, 2019] extracted the visual features through GNN to enhance the input of the BiLSTM-CRF model, which was proved to be effective. Different from the fully-connected or hand-crafted graph, the PICK [Yu *et al.*, 2020] predicted the connections between nodes through graph learning [Jiang *et al.*, 2019a], which also boosts the results. These methods used the GNN to encode text embeddings given visually rich context to learn the key-value relationship implicitly. It is difficult to ensure that models can learn it well. However, our method explicitly learns key-value matching which makes full use of edge features for relevancy evaluation.

3 Methodology

The framework of the proposed MatchVIE is presented in Figure 2. It consists of a multi-feature extraction backbone and two specific branches of relevancy evaluation and entity recognition, respectively. The multi-feature extraction backbone considers features (e.g. position, image, and semantics) all together. Then the relevancy evaluation branch is based on a graph module to model the overall layout information of documents and obtain the key-value matching probability. Meanwhile, to solve the sequence labels of standalone text segments, we design an entity recognition branch.

3.1 Multi-feature Extraction Backbone

To explore the features of multimodal, we propose an effective way to consider position features, visual features and textual features. We consider the input to be a set of text segments. As for textual features, given the i-th text segment $T_i = (t_1^i, t_2^i, t_3^i, ..., t_L^i)$, where L is the number of tokens in it, and t_m is the m-th token embedding in the text segment. Inspired by [Devlin *et al.*, 2019], we apply the pre-train model to initialize the parameters. As for position embedding, different from the position embedding that represents

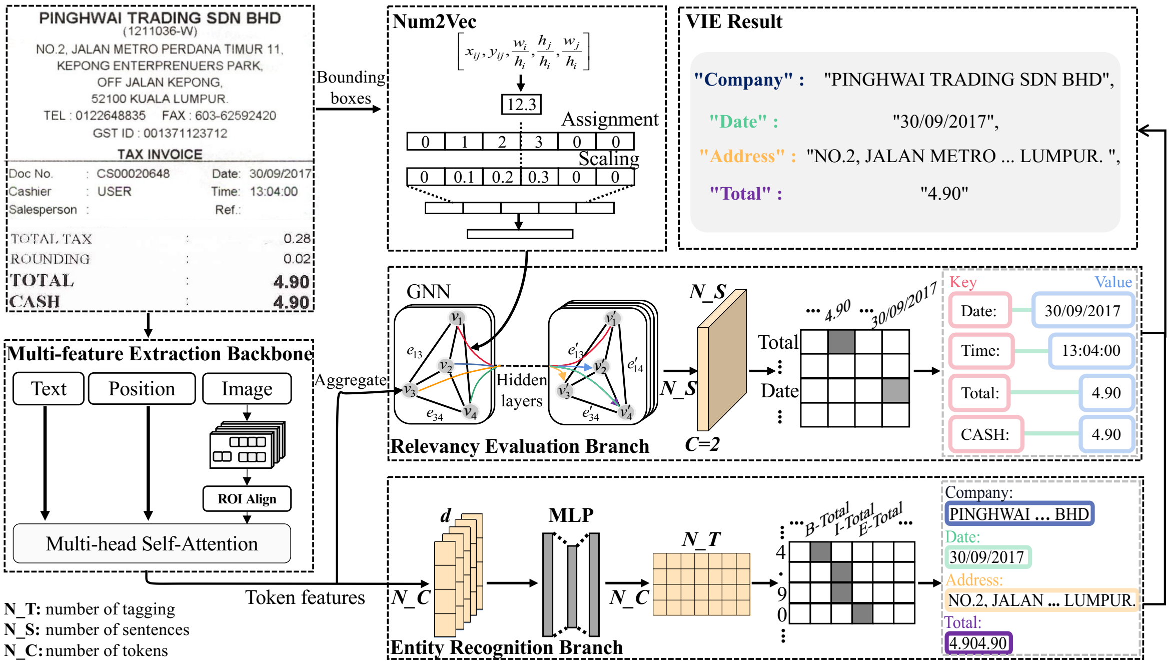


Figure 2: Overall framework of MatchVIE. The relevancy evaluation branch predicts the key-value relationship between entities. The entity recognition branch mainly determines the categories of standalone entities. The entity recognition branch is difficult to distinguish numeric categories which are similar in visual and semantic, such as the '4.90' in purple.

the word sequence, we use the spatial position of each token in the document. In detail, given a position embedding $P_i = (p_1, p_2, p_3, ..., p_L)$, the position embedding p_m for m-th token is obtained by linearly transforming its bounding box coordinates to a d-dim vector.

As for image visual embedding, previous studies [Yu *et al.*, 2020] usually applied a CNN for catching morphology information from each cropped image. This is unable to capture global visual information because of isolating each text segment. Therefore, we attempt to extract the features from the whole image by ResNet [Boroumand *et al.*, 2018]. Then, the ROI-Align [He *et al.*, 2017] is used to obtain the Region of Interest (ROI) according to the coordinate of tokens. Similarly, the visual embedding of the i-th text segment is denoted by $I_i = (i_1, i_2, i_3, ..., i_L)$. The m-th token visual embedding is given as follows:

$$I_m = ROIAlign(ConvNet(X)), \tag{1}$$

where X is the inputted image. In order to capture local and non-local features, we use the self-attention mechanism to obtain the context features $\hat{C}$. Afterwards, we obtain the Q,K, V in the scaled dot-product attention by combining the position features, visual features and textual features together. In detail, the context feature $\hat{C}$ is obtained by,

$$\begin{aligned} \hat{C} &= MultiHead(Q, K, V) \\ &= [head_1, head_1, ..., head_n]W^d, \end{aligned} \tag{2}$$

$$Q, K, V = LayerNorm(Linear(I) + P + T), \tag{3}$$

where T, P and I is the textual embedding, position embedding and visual embedding, respectively. The d is the embedding dimension intended to be 768.

3.2 Relevancy Evaluation Branch

To represent the relevancy between entities, the layout information such as the distance between text segments is an important clue as demonstrated in [Cheng *et al.*, 2020]. We construct a fully-connected graph for the document, which means each node is connected to each other. Given a graph $G = (V, E)$, $v_i \in V(i = 1, 2, ..., N)$ denotes one of the N nodes (text segments) in a graph document, and $e_{ij} \in E$ represents the edge between node v_i and v_j.

The initial node features are obtained by aggregating the context feature $\hat{C}$ in units of text segments. The initial edge features are expected to represent the relationships in vision between text segments. Therefore, we encode the visual information into the edges. Previously, [Liu *et al.*, 2019] defined edge features as,

$$\mathbf{e}_{ij} = [x_{ij}, y_{ij}, \frac{w_i}{h_i}, \frac{h_j}{h_i}, \frac{w_j}{h_i}], \tag{4}$$

where x_{ij}, y_{ij} represents the relative position between node v_i and v_j; w_i/h_i denotes the aspect ratio of node v_i. However, we find that the encoded numeric values are very unstable because of the diversity of distance and shape of different text segments. To address this issue, we propose a simple yet effective operation, namely Num2Vec to process each item. As shown in Figure 2, for a numeric value, we provide a fixed-length of 8 arrays to hold each digit position. The first half of the array corresponds to an integral part, and the rest corresponds to the fractional part. Then, we scale these digits by a factor of 0.1. Consequently, the encoded values are constrained to the range of $[-0.9, +0.9]$, which can effectively reduce the fluctuation range of the data.

In the process of features update in the GNN, these two types of features are mixed together for later predictions. We follow the approach adopted by [Liu *et al.*, 2019] who defined a triplet $(\mathbf{v}_i, \mathbf{e}_{ij}, \mathbf{v}_j)$ for features updating. The triplet feature is linearly transformed by a learnable weight $\mathbf{W}_g$, which generates an intermediate feature $\mathbf{g}_{ij}$,

$$\mathbf{g}_{ij} = \mathbf{W}_g[\mathbf{v}_i || \mathbf{e}_{ij} || \mathbf{v}_j]. \tag{5}$$

Then, we apply $\mathbf{g}_{ij}$ to update the edge features through another linear transformation (applying a nonlinearity, σ):

$$\mathbf{e}_{ij}{}' = \sigma(\mathbf{W}_e \mathbf{g}_{ij}), \tag{6}$$

where the $\mathbf{g}_{ij}$ is also used to evaluate the attention coefficients and the new node features. The $\mathbf{e}_{ij}{}'$ is the intermediate update state of edge features. We also used multi-head attention [Veličković *et al.*, 2017] to enhance the feature learning.

Relevancy Evaluation. The combination of text segments can be enumerated by N^2 edges. We model the relevancy evaluation as a binary classification problem. If two text segments form a key-value pair, the combination is treated as positive; otherwise, it is negative. In detail, we feed the last edge features $\mathbf{e}_{ij}$ to another multi-layer perceptron (MLP) which predicts two logits for each edge. After performing softmax activation on the logits, we acquire a matching probability regarded as relevancy between two text segments. When training this branch, it is notable that the matching matrix of size $N \times N$ is very sparse for the correct combinations. The loss of positive samples is easily overwhelmed by the negative ones. To solve this problem, the loss function for the relevancy evaluation branch is designed as the focal loss [Lin *et al.*, 2017] which can automatically adjust weight coefficients based on the difficulty of samples. The procedure is as follows:

$$\mathbf{p}_{ij} = Softmax(\mathrm{MLP}(\mathbf{e}_{ij})), \tag{7}$$

$$\mathcal{L}_{Re}(p, y^*) = \begin{cases} -\alpha(1 - p'_{ij})^{\gamma} log(p'_{ij}), & y^*_{ij} = 1, \\ -(1-\alpha) p'^{\gamma}_{ij} log(1 - p'_{ij}), & otherwise, \end{cases} \tag{8}$$

where $\mathbf{p}_{ij}$ is the predicted probability vector and the $p^{'}_{ij}$ represents the probability for the positive class. The y^*_{ij} is the label for edge e_{ij}. The γ is a focusing parameter intended to be 2. The α is used to balance the positive and negative classes and we set it to 0.75 in our experiments.

3.3 Entity Recognition Branch

Some standalone text segments are usually easy to distinguish by semantics, but their attributes couldn't be determined without key-value pairs. To address this issue, we design the entity recognition branch. Note that this branch is oriented to standalone text segments. The implementation for this branch is not unique, here we follow common sequence labeling methods. We feed the context feature $\hat{C}$ to a MLP, projecting the output to the dimension of BIOES tagging [SANG, 1999] space. For each token, we perform element-wise sigmoid activation on the output vector to yield the predicted probability of each class,

$$P_{entity} = Softmax(\mathrm{MLP}(\hat{C})). \tag{9}$$

Then a CRF layer is applied to learn the semantics constraints in an entity sequence. For CRF training, we minimize the negative log-likelihood estimation of the correct entity sequence and calculate the loss $\mathcal{L}_{entity}$.

3.4 Two-branch Merge Strategy

During training, the proposed two branches can be trained and the losses are generated from two parts,

$$\mathcal{L} = \lambda_{entity}\mathcal{L}_{entity} + \lambda_{Re}\mathcal{L}_{Re}, \tag{10}$$

where hyper-parameters λ_{entity} and λ_{Re} control the trade-off between losses, we set them to 1.0. During inference, It is worth to note that we give priority to the classification results of the relevancy evaluation branch. In relevancy evaluation branch, if the matching probability is higher than a threshold (0.5), these two text segments are regarded as a key-value pair. For remaining texts which are considered as standalone texts, prediction results of the entity recognition branch are adopted. The attribute (category) of a value can be determined according to its corresponding key. A simple approach is to perform lookup on a mapping table that contains all possible keys for a certain category. The mapping table needs to be created in advance based on training set. It is not flexible to customize all the categories. Therefore, we propose another way to determine the category based on the semantic similarity. The details are as follows:

$$\mathbf{Z} = \min_{\mathrm{j} \in \{\mathrm{C}\}} \|\mathrm{BiLSTM}(\mathbf{x}^{\mathrm{i}}_{1:\mathrm{m}}; \mathbf{\Theta}') - \mathrm{BiLSTM}(\mathbf{y}^{\mathrm{j}}_{1:\mathrm{n}}; \mathbf{\Theta}')\|_2, \tag{11}$$

where $\mathbf{x}^i_{1:m}$ are word embedding for each token in the key. Similarly, the $\mathbf{y}^j_{1:n}$ are those for the $category$ name. We use the pre-trained models to initialize the word embedding features (semantic features) of texts. To acquire the semantic representation, we input the key and $category$ name to an off-the-shelf BiLSTM model to get the context encoding over all tokens. The $\mathbf{\Theta}'$ is the learnable weight of BiLSTM which is fixed during inference. We then compute the L2 distance for the semantic feature vectors of the key and $category$ name. The one with the minimum distance is selected as the category of the key-value pair.

4 Experiments

4.1 Datasets

We conduct experiments on three real-world public datasets, whose statistics are given in Table 1. Note that the layout of the three datasets is variable.

Dataset	Training	Testing	Entities	K-V Ratio(%)
FUNSD	149	50	4	74.69
EPHOIE	1183	311	10	85.62
SROIE	626	347	4	54.33

Table 1: Statistics of datasets used in this paper, including the division of training/testing sets, the categories of named entities, and the proportion of key-value pairs.

FUNSD[Jaume *et al.*, 2019] is a public dataset of 199 fully annotated forms, which has 4 entities to extract (i.e., Question, Answer, Header and Other).

Module	FUNSD	EPHOIE	SROIE
	F1(%)	F1(%)	F1(%)
MatchVIE(Ours)	**81.33**	**96.87**	**96.57**
(-)Focal Loss	80.66	96.28	95.21
(-)K-V matching	76.47	92.19	93.23
(-)Num2Vec	74.25	90.31	91.31

Table 2: Ablation study (F1-score) of the proposed model on the three datasets.

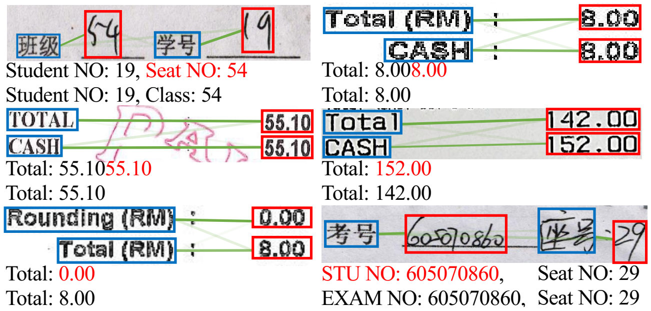


Figure 3: Comparison between the prediction results of MatchVIE (**bottom row**) and sequence labeling methods (**top row**). Red blocks: $value$, blue blocks: key, red font: error prediction, green line depth: matching confidence level.

EPHOIE[Wang *et al.*, 2021] is a public dataset that consists of 1,494 images of Chinese examination paper head which has 10 entities to extract (i.e., Subject, Name, Class).

SROIE[Huang *et al.*, 2019] is a public dataset that contains 973 receipts in total. Each receipt is labeled with 4 types of entities (i.e., Company, Address, Date, and Total).

4.2 Implementation Details

The model is trained from using the Adam optimizer with a learning rate of 0.0005 to minimize the $\mathcal{L}_{entity}$ and $\mathcal{L}_{Re}$ jointly during the training. The feature extractor for catching image features is implemented by ResNet-34 [Boroumand *et al.*, 2018]. We apply the BIOES tagging scheme for entity recognition branch. In relevancy evaluation branch, we set the number of graph convolution layers to 2 and 8 heads for the multi-head attention. We use the text-lines which have already been annotated in the datasets as the text segments.

4.3 Ablation Study

As shown in Table 2, we analyze the effect of each component in MatchVIE, including the focal loss, K-V matching, Num2Vec. We set up the changes in model accuracy when these three modules are not considered. Without focus loss, the relevancy evaluation branch cannot effectively overcome the problem that matching matrix is very sparse, especially for SROIE dataset because of the little proportion of key-value pair. It can be seen that the relevancy evaluation branch (K-V matching) can improve the accuracy by a large margin. Without it, the performance of MatchVIE will decrease 4.09 % in EPHOIE dataset beacause there is 85.62% K-V ratio in it. In order to further verify the effectiveness of the relevancy evaluation branch, we give some prediction results of the MatchVIE model on whether removing the relevancy evaluation branch or not. Note that when the relevancy evaluation branch is removed, our method becomes sequence labeling methods by relying on the prediction of entity recognition branch. From Figure 3, it can be seen that sequence labeling methods are not effective enough for distinguishing numeric semantic categories or some ambiguous texts. On the contrary, our MatchVIE can effectively distinguish these categories by introducing the correlation between named entities. Besides, after employing Num2Vec, the model can achieve a more stable results with extra accuracy improvements. In addition, we collect training loss and plot the loss curves. From Figure 4, it can be seen that Num2Vec can help the model converge more smoothly.

Setting	Meathods	FUNSD	EPHOIE	SROIE
		F1(%)	F1(%)	F1(%)
1)	Lookup table	80.44	**96.96**	96.27
Ours	Semantic similarity	**81.33**	96.87	**96.57**

Table 3: Two methods of mapping keys to certain categories.

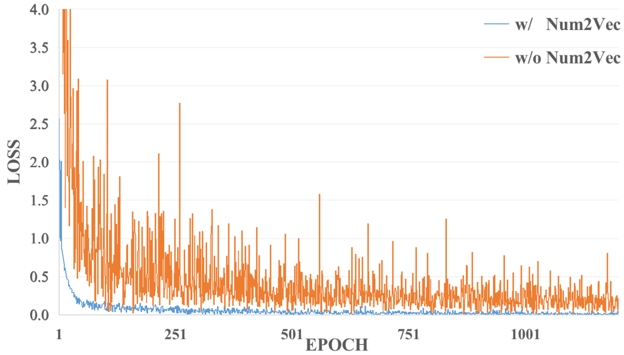


Figure 4: The yellow polyline indicates loss without Num2Vec and the blue polyline indicates the loss with Num2Vec. It can be seen from that using Num2Vec can make training converge smoothly.

4.4 Explorations of Network Architecture

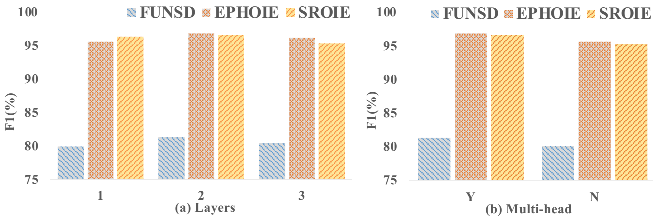
For the relevancy evaluation branch, we evaluate network architecture with different structures.

Comparisons of Different Mapping Methods. We try two methods to map a key to a certain category. One is the lookup table given all possible keys for each category (e.g, the lookup table for the keys of the category 'Total' can be 'Total Sales', 'Total', 'Total Amount', 'Total (RM)'), and the other is based on semantic similarity, detailed in section 3.4. As shown in Table 3, the lookup method is slightly better than semantics based method in EPHOIE dataset. However, the semantics based methods can flexibly customize categories and they don't need to build the lookup table in advance. Either method can be selected based on the actual situation.

Network Architecture with Different Settings. For the GNN in relevancy evaluation branch, we explore the appropriate network architecture with different settings. By default, we use 8 heads for the multi-head attention. As for the layer number, the performance of two layers is better than that of

Method	**Entities**										
	Subject	Test Time	Name	School	EXAM NO	Seat NO	Class	STU NO	Grade	Score	**F1(%)**
LSTM-CRF[Lample *et al.*, 2016]	98.51	**100.0**	98.87	98.80	75.86	72.73	94.04	84.44	98.18	69.57	89.10
([Liu *et al.*, 2019])	98.18	**100.0**	99.52	**100.0**	88.17	86.00	97.39	80.00	94.44	81.82	92.55
GraphIE ([Qian *et al.*, 2019])	94.00	**100.0**	95.84	97.06	82.19	84.44	93.07	85.33	94.44	76.19	90.26
TRIE [Zhang *et al.*, 2020]	98.79	**100.0**	99.46	99.64	88.64	85.92	97.94	84.32	97.02	80.39	93.21
MatchVIE(Ours)	**99.78**	**100.0**	**99.88**	98.57	**94.21**	**93.48**	**99.54**	**92.44**	**98.35**	**92.45**	**96.87**

Table 4: Experiment results on EPHOIE datasets. Standard F1-score (F1) are employed as evaluation metrics. LayoutLM only has pre-training models on English data and doesn't work on EPHOIE which is in Chinese. NO: Number, STU:Student, EXAM: Examination.



(a) Layers

(b) Multi-head

Figure 5: Impacts of layers and heads in the relevancy evaluation branch. The impacts of number of layers (left). The impacts of whether using muti-head or not (right).

Method	**FUNSD**	**SROIE**
	F1(%)	**F1(%)**
LSTM-CRF[Lample *et al.*, 2016]	62.13	90.85
GraphIE [Qian *et al.*, 2019]	72.12	94.46
([Liu *et al.*, 2019])	72.43	95.10
LayoutLM [Xu *et al.*, 2020] †	79.27	95.24
TRIE [Zhang *et al.*, 2020]	78.86	96.18
MatchVIE(Ours)	**81.33**	**96.57**

Table 5: Experiment results on FUNSD dataset and SROIE dataset. † indicates the results is reported in [Xu et al. 2019]

one layer, as shown in Figure 5 (a). However, when three layers are used, the additional layer does not lead an evident gain. The results in Figure 5 (b) reveals that the multi-head attention can improve the accuracy to some extent.

4.5 Comparisons with State-of-the-arts

We compare our proposed model with the following several methods: (1) GraphIE [Qian *et al.*, 2019] is a VIE framework that learns local and non-local contextual representations by graph structure. (2) Liu et al. [Liu *et al.*, 2019] introduces a GNN model to combine global layout information into BiLSTM-CRF. (3) LayoutLM [Xu *et al.*, 2020] is an effective pre-training method with text and layout information. (4) TRIE [Zhang *et al.*, 2020] is a end-to-end method that combines visual and textual features for VIE. We re-implement them based on the original papers or source codes.

Results on EPHOIE Dataset. The layout of images in EPHOIE dataset is complex and there are multiple types of entities, which are easy to confuse. Most of these entities are numbers and do not contain semantics (e.g, 'EXAM NO', 'Seat NO', 'STU NO' and 'Score'). As shown in Table 4, to better illustrate the complementarity of our proposed MatchVIE with other methods, we give the F1-Score of all categories. Sequence labeling methods conduct sequence labeling on the documents where text segments are concatenated from left to right and from top to bottom. This causes the determination of these numeric semantic categories to rely on the pre-organized order. Mainly owing to the strong correlation between entities, our MatchVIE can bypass the recognitions to various semantics, and simply focuses on the strong relevancy between entities. It achieves best results and outperforms the state-of-the-art methods by significant margins (3.66% than TRIE), especially in numeric semantic categories or some ambiguous texts.

Results on FUNSD Dataset. As shown in Table 5, some multimodal methods, such as GraphIE and Liu's methods, achieve high performance by introducing layout information. Meantime, LayoutLM achieves a better performance of 79.27% when using the text, layout and image information at the same time. Our method outperforms the state-of-the-art results by 2.06% due to the introduction of entity relationship.

Results on SROIE Dataset. The results of experiments on SROIE datasets are shown in Table 5. Our MatchVIE model achieve a F1-score of 96.57%, which is better than the first place in SOTA methods. In this dataset, the location of these two categories of 'Total' and 'Date' are more flexible compared to other categories, and they are easily confused with other categories, as shown in Figure 3. Our method can effectively distinguish these two categories by combining neighbor relationships. Although the relatively little proportion of key-value (54.33%), our method has achieved competitive results.

5 Conclusion

In this paper, we propose a novel MatchVIE model to extract the information from document images. Through the key-value matching, the proposed MatchVIE can bypass the recognition to various semantics, and simply focuses on the strong relevancy between keys and values which has been guaranteed by the document layout or semantic information. Comprehensive experiments show the benefit of learning the key-value matching relationship explicitly. We plan on extending this work to transfer the layout information between document images to achieve One-shot learning.

Acknowledgments

This research is supported in part by NSFC (Grant No.: 61936003, 61771199), GD-NSF (no.2017A030312006).

References

[Boroş *et al.*, 2020] Emanuela Boroş, Verónica Romero, Martin Maarand, and Zenklová. A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In *ICFHR*, pages 79–84, 2020.

[Boroumand *et al.*, 2018] Mehdi Boroumand, Mo Chen, and Jessica Fridrich. Deep residual network for steganalysis of digital images. *IEEE Trans. Inf.*, 14(5):1181–1193, 2018.

[Carbonell *et al.*, 2020] Manuel Carbonell, Alicia Fornés, Mauricio Villegas, and Josep Lladós. A neural model for text localization, transcription and named entity recognition in full pages. *Pattern Recognit Lett*, 2020.

[Cheng *et al.*, 2020] Mengli Cheng, Minghui Qiu, Xing Shi, Jun Huang, and Wei Lin. One-shot text field labeling using attention and belief propagation for structure information extraction. In *ACMmm*, pages 340–348, 2020.

[Denk and Reisswig, 2019] Timo I Denk and Christian Reisswig. BERTgrid: Contextualized embedding for 2d document representation and understanding. In *NIPS*, 2019.

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

[Gal *et al.*, 2020] Rinon Gal, Shai Ardazi, and Roy Shilkrot. Cardinal graph convolution framework for document information extraction. In *ACMmm*, pages 1–11, 2020.

[Gui *et al.*, 2019] Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuan-Jing Huang. A lexicon-based graph neural network for chinese NER. In *EMNLP*, pages 1039–1049, 2019.

[Guo *et al.*, 2019] He Guo, Xiameng Qin, and Liu. EATEN: Entity-aware attention for single shot visual text extraction. In *ICDAR*, pages 254–259, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask RCNN. In *CVPR*, pages 2961–2969, 2017.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[Huang *et al.*, 2019] Z Huang, K Chen, J He, X Bai, D Karatzas, S Lu, and CV Jawahar. ICDAR2019 competition on scanned receipt ocr and information extraction. In *ICDAR*, pages 1516–1520, 2019.

[Hwang *et al.*, 2019] Wonseok Hwang, Seonghyeon Kim, Minjoon Seo, Jinyeong Yim, Seunghyun Park, Sungrae Park, Junyeop Lee, Bado Lee, and Hwalsuk Lee. Post-ocr parsing: building simple and robust parser via bio tagging. In *NIPS*, 2019.

[Jaume *et al.*, 2019] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDARW*, volume 2, pages 1–6, 2019.

[Jiang *et al.*, 2019a] Bo Jiang, Ziyan Zhang, Doudou Lin, Jin Tang, and Bin Luo. Semi-supervised learning with graph learning-convolutional networks. In *CVPR*, pages 11313–11320, 2019.

[Jiang *et al.*, 2019b] Zhaohui Jiang, Zheng Huang, Yunrui Lian, Jie Guo, and Weidong Qiu. Integrating coordinates with context for information extraction in document images. In *ICDAR*, pages 363–368, 2019.

[Katti *et al.*, 2018] Anoop Raveendra Katti, Christian Reisswig, and Guder. Chargrid: Towards understanding 2d documents. In *EMNLP*, pages 4459–4469, 2018.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016.

[Lample *et al.*, 2016] Guillaume Lample, Miguel Ballesteros, and Subramanian. Neural architectures for named entity recognition. In *NAACL*, pages 260–270, 2016.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

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

[Qian *et al.*, 2019] Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. GraphIE: A graph-based framework for information extraction. In *NAACL*, 2019.

[SANG, 1999] EFTK SANG. Representing text chunks. In *EACL*, pages 173–179, 1999.

[Veličković *et al.*, 2017] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2017.

[Wang *et al.*, 2021] J Wang, C Liu, L Jin, and G Tang. Towards robust visual information extraction in real world: New dataset and novel solution. In *AAAI*, 2021.

[Xu *et al.*, 2020] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. *KDD*, 2020.

[Yu *et al.*, 2020] Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. PICK: Processing key information extraction from documents using improved graph learning-convolutional networks. *ICPR*, 2020.

[Zhang *et al.*, 2020] Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. TRIE: End-to-end text reading and information extraction for document understanding. *ACMmm*, 2020.