CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor

Xiaohui Zhao, Zhuo Wu, and Xiaoguang Wang
New IT Accenture
xiaohui.zhao@accenture.com zhuo.wu@accenture.com danny.x.wang@accenture.com

Abstract

Extracting key information from documents, such as receipts or invoices, and preserving the interested texts to structured data is crucial in the document-intensive streamline processes of office automation in areas that includes but not limited to accounting, financial, and taxation areas. Large proportion of published works attempt to tackle the problem by exploring the semantic context in text sequences based on the Named Entity Recognition (NER) method in the NLP field. In this paper, we propose to combine the effective information from both semantic meaning and spatial distribution of texts in documents. Specifically, our proposed model, Convolutional Universal Text Information Extractor (CUTIE), applies convolutional neural networks on the gridded texts where texts are embedded as features with semantical connotations. We further explore the effect of employing different structures of convolutional neural network and propose a faster and portable structure. We demonstrate the effectiveness of the proposed method on a dataset with up to 5000 *labelled receipts, without any pretraining or post-processing, achieving state of the art performance that is higher than BERT but with only 1/10 parameters and much higher than the other RNN based methods.*

1. Introduction

Implementing Scanned receipts OCR and information extraction (SROIE) is of great benefit to services and applications such as efficient archiving, compliance check, and fast indexing in the document-intensive streamline processes of office automation in areas that includes but not limit to accounting, financial, and taxation areas. There are two specific tasks involved in SROIE: receipt OCR and key information extraction. In this work, we focus on the second task that is rare in published research. In fact, key information extraction faces big chanllenges, where different types of document structures and vast number of potential interested key words introduces great difficulties. Although the commonly used rule-based method can be implemented with carefully designed expert rules, it can only work on specific type of documents and takes no lesser effort to adapt to new type of documents. Therefore, it is desirable to have a learning based key information extraction method with limited requirement of human resources and solely employing the deep learning technique without designing expert rules for any specific type of documents.

History about published learning based methods goes here. The majority of the published learning based works are based on the Named Entity Recognition (NER) research in the NLP field that is not originally designed to solve the key information extraction problem in SROIE, which align text words in the original document as a long paragraph in line-based rule. However, the real world documents, such as receipts and invoices, present with various styles of layouts that were designed for different scenarios or from different enterprise entities. The order or word-to-word distance of the texts in the line-based aligned long paragraph tend to vary greatly due to layout variations, which is difficult to be handled with the natural language paragraph oriented methods, one typical example is illustrated in Fig. 2.

In this work, attempting to involve the spatial information into the key information extraction process, we propose to tackle this problem by using the CNN based network structure and involve the semantic features in a properly designed fashion. In particular, our proposed model, called Convolutional Universal Text Information Extractor (CUTIE), tackles the key information extraction problem by applying convolutional deep learning model on the gridded texts, as illustrated in Fig. 1. The gridded texts is formed with the proposed grid positional mapping method, where the grid is generated with the principle that is preserving text's relative spatial relationship in the original scanned document image. The rich semantic information is encoded from the gridded texts at the very beginning stage of the convolutional neural network with a word embedding layer. The CUTIE allows for simutaneously looking into both semantical information and spatial information of the texts in

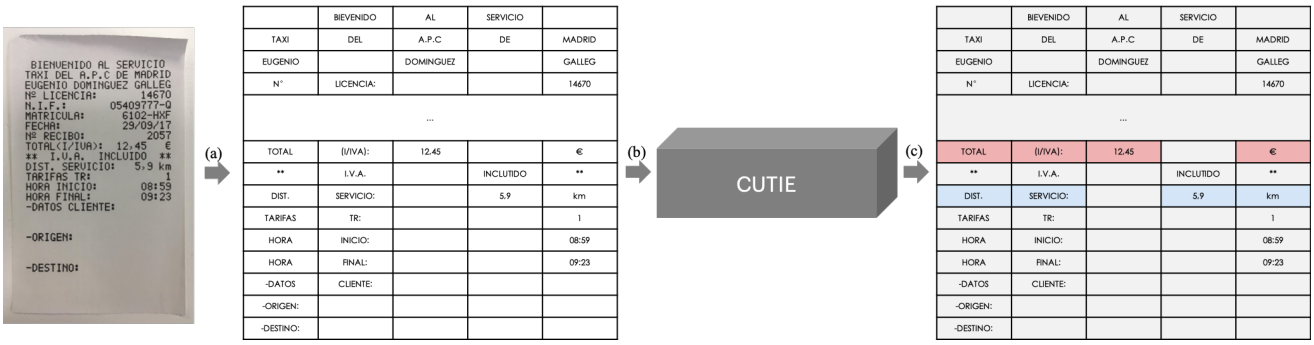

Figure 1. Framework of the proposed method, (a) positional map the scanned document image to a grid with text's relative spatial relationship preserved, (b) feed the generated grid into the CNN for extracting key information, (c) reverse map the extracted key information for visual reference.

the scanned document image and can reach a new state of the art result for key information extraction, which outperforms BERT model but without demanding pretrain on a huge text dataset [5, 9].

2. Related Work

3. Methods

In this section, we introduce the method proposed for creating grid data for model training. We then present the network architectures that capture long distance information and avoid information loss in the convolutional neural networks that have striding or pooling processes.

3.1. Grid Positional Mapping

To generate input grid data for the convolutional neural network, the scanned document image are processed by an OCR engine to acquire the texts and their absolute / relative positions. Let the scanned document image be of shape (w, h), the minimum bounding box around the i-th interested text s_i be b_i that is restricted by two corner coordinates, where the upper-left corner coordinate in the scanned document be (x^i_{left}, y^i_{top}) and the bottom right of the bounding box be $(x^i_{right}, y^i_{bottom})$. To avoid the affects from overlapped bounding boxes and reveal the actual relative position among texts, we calculate the center point (c^i_x, c^i_y) of the bounding boxes as the reference position. It is not hard to find that involving pre-processes that combine texts into meaningful entities will benefit the grid positional mapping process. However, this is not the major purpose of this paper and we leave it to future researches. In this paper, we tokenize the text words with a greedy longest-match-first algorithm using a pre-defined dictionary [6].

Let the grid positional mapping process be G and the target grid size be (c_{g_m}, r_{g_m}). To generate the grid data, the goal of G is to map the texts from the original scanned document image to the target grid, such that the mapped grid preserves the original spatial relationship among texts yet more suitable to be used as the input for the convolutional neural network. The mapping position of texts in the grid is calculated as

$$c^i_x = c_{g_m} \frac{x_{left} + \frac{(x_{right} - x_{left})}{2}}{w} \tag{1}$$

$$r^i_y = r_{g_m} \frac{y_{top} + \frac{(y_{bottom} - y_{top})}{2}}{h} \tag{2}$$

For tokenized texts, the bounding box is horizontally divided into multiple boxes and their row and col reference positions are calculated using the same criteria as Equ. 1 and Equ. 2, separately. Furthermore, to enhance the capability of CUTIE to better handle documents with different layouts, we augment the grid data to shapes with different rows and columns by random sampling a Gaussian distribution for with probability

$$p_c(k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(k - c_{g_t})^2}{2\sigma^2}} \tag{3}$$

$$p_r(k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(k - r_{g_t})^2}{2\sigma^2}} \tag{4}$$

where c_{g_t} is the mean center of the target augment gird size, r_{g_t} is the mean center of the target augment grid size, and σ is the standard deviation.

3.2. CUTIE Model

Through matching the output of CUTIE with the labelled grid data, the model learns to generate the label for each text in the grid input via exploring both the spatial and semantic features. For that reason, the task of CUTIE bears resemblance to the semantic segmentation task in the computer vision field but with more sparse data distributions. Specifically, the mapped grid contains scattered data points (text tokens) in contrast to the images bespread with pixels. The grid positional mapped key texts are either close

to or distant to each other due to different types of document layouts. Therefore, incorporating multi-scale context processing ability benefits the network.

In fact, several methods have been proposed in the semantic segmentation field to capture multi-scale contexts in the input data. The methods of image pyramid and the encoder-decoder structure both aim at exploiting multi-scale information. The interested objects from different scales become prominent in the former networks by using multiple scaled input data to gather multi-scale features. The later networks shrink feature maps to enlarge receptive fields and reduce computation burdens, and then capture finer details by gradually recovering the spatial information from lower layer features. However, spatial resolution is reduced in the encoding process and the decoding process exploits only high resolution but low level features to recover the spatial resolution, the consecutive striding encoding process decimates detail information [8]. Moreover, the encoding and decoding process applies shape restricts to the grid shape augment process as introduced in Section 3.1.

Instead, the field of view of filters can also be effectively enlarged and multi-scale contexts can be captured by combining multi-resolution features [8] or by applying atrous convolution [1, 2, 3, 4]. To capture long distance connection and avoid potential information loss in the encoding process, we propose two different network architectures and compare their performance in Section 4. Specifically, proposed CUTIE-A is a high capacity convolutional neural network that fuses multi-resolution features without losing high-resolution features, proposed CUTIE-B is a convolutional network with atrous convolution for enlarging field of view and Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale contexts.

Both CUTIE-A and CUITE-B conducts semantical meaning encoding process with a word embedding layer in the very beginning stage. The cross entropy loss function is applied to compare the predicted token class grid and the ground truth grid.

3.2.1 CUTIE-A

CUTIE-A avoids information loss in the encoding process while taking advantage of encoders by combining encoding results to the maintained high-resolution representations through the entire convolutional process. Similar to HRNet proposed in [8], a high-resolution network without striding is employed as the backbone network and several high-to-low resolution sub networks are gradually added and connected to the backbone major network. During the connecting process of the major network and sub networks, multi-scale features are fused to generate rich representations.

3.2.2 CUTIE-B

CUTIE-B is constructed with a single backbone network but employs atrous convolution to capture long distance connections. For atrous convolution, let the input feature map be m, filter be w and output be n, for each position i, atrous convolution is applied over the input feature map m as

$$n[i] = \sum_k m[i + r \cdot k]w[k] \tag{5}$$

where r is the atrous rate that indicates the sampling stride of the input signal, which is implemented as convolving the input feature with upsampled filters by inserting $r - 1$ zeros between two consecutive filter values along each spatial dimension. Standard convolution is a special case of atrous convolution with $r = 1$ [2].

4. Experiments

The proposed method is evaluated on a dataset with 3 types of scanned document images, which contain 8 key information classes and 1 don't care class. For each specific key information class, multiple tokens can be included. The overall performance is refered as strict average precision (AP) and measured in terms of per-class accuracy across the 9 classes, where one class is determined as correct only when every single token in the class is correct. To achieve deeper analysis of the performance of the proposed method, we propose to use one more criteria, soft average precision (softAP), where the prediction of a key information class is determined as correct as if positive ground truths are correctly predicted even if some false positive are included in the final prediction. SoftAP is important since it indicates model's capability of extracting correct key information with tolerance of incorporating certain false positives. In fact, post processings can be employed there to eliminate the false positives. Therefore, joint analysis of AP and softAP provides a better understanding of the model performance.

We compare the performance of the proposed method with two state of the art methods CloudScan [7] and BERT [5]. For comparison, the Cloud Scan model for SROIE is trained from scratch but with several expert designed features as describbed in [7]. The BERT model for SROIE is transform learned with the Google released base model that is pre-trained on a huge dataset with $3,300M$ words [5, 6].

We use a learning rate of $1e{-}3$ with Adam optimizer and step decay learning strategy. The learning rate is dropped to $1e{-}4$ and $1e{-}5$ on the 15000-th and 30000-th steps, respectively. The training is terminated within 40000 steps with batch size of 32. Our model is trained end-to-end without piecewise pretraining of any component. The embedding size is 128, target augmentation shape is 64 for both row

Figure 2. Example of scanned taxi receipt images. We provide two colored rectangles to help readers find the key information about distance of travel and total amount with blue and red, respectively. Note the different types of spatial layouts and key information texts in these receipt images.

and column. The dataset is split as training set and test set with ratio of $7:3$.

4.1. Dataset

The dataset contains 4000 annotated scanned spanish receipt documents, including taxi receipts and meal entertainment (ME) receipts, with 9 different key information classes, as reported in Table 1. We generate the texts and bounding boxes with Google's OCR API. Each text and their bounding box is manually labelled as one of the 9 different classes: 'DontCare', 'VendorName', 'VendorTaxID', 'InvoiceDate', 'InvoiceNumber', 'ExpenseAmount', 'BaseAmount', 'TaxAmount', and 'TaxRate'. We then employ the tokenizer introduced in Section 3.1 to segment texts into minimum token units, where text bounding boxes are also segmented accordingly.

In fact, the dataset is much difficult than the neat scanned document images, since various layouts of receipts were captured in different scenarios using ordinary mobile phone camera. Examples of the scanned document images in our dataset are illustrated in Figure 2, note that the colored rectangles is only for visual reference, the actual labelled data is in token-level rather than the line-level that is shown in the figure.

4.2. Results on Validation Set

We report results of our method and compare with other state of the art methods in Table 2. Our big network CUTIE-A achieves 100 score on meals entertainment receipt and 100 score on taxi receipt, outperforming other methods. Compared to CloudScan, our big network CUTIE-A improves AP by 100 on meals entertainment receipt and 100 on taxi receipt, our small network CUTIE-B improves AP by 100 points on meals entertainment receipt and 100 on taxi receipt. Compared to BERT, where model is pre-trained on a dataset with $3,300M$ words and transfer learned on our dataset, our big network CUTIE-B improves AP by 100 100 while using only $1/2$ parameters, our small network CUTIE-B improves AP by 100 100 but with much less complexity and smaller model size with only $1/10$ parameters. Examples of inference results are illustrated in Figure 3.

Table 1. Number of labelled receipt document images and key information classes of different types of documents in the labelled dataset.

	Training Set	Test Set	#classes
ME	1109	475	9
Taxi	2514	1077	6

Table 2. Performance comparison on 3 types of documents. (AP/softAP)

Method	#Params	ME	Taxi	Hotel
CloudScan	-	82 / -	64 / -	
BERT	110M	80.3 / -	87.0 / -	79.4 / -
CUTIE-A	67M	82.5 /	/	/
CUTIE-B	14M	/	/	

We also provides softAP results for CUTIE-A and CUTIE-B in Table 2, where both the softAP of CUTIE-A and CUTIE-B exceeds AP by a large margin. Typical examples of receipts with low AP but high softAP score is shown in Figure 4. Most of the false positive cases occor in the 'VenderName' class, where names tend to greatly vary and leads to difficulty in model inference. It is not hard to find that, however, these false positives can be easily avoided by appending a dictionary based post processor to the key information extractor. One rare false posive case in the 4-th image is a 'L' letter being mis-recognized as '1' by the employed OCR engine. Although the letter shows distinct appearance with the other digits, it is misinterpreted as one part of the 'BaseAmount' class due to its close spatial location to the digits and it-self being a digit. Although this is one and only special case in our test set, it still suggests that incorporating image-level information may further boost the inference accuracy inspite of the already involved semantical and spatial features.

Furthermore, we find that some cases are determained as not right while the ground truth were wrongly labelled by the human labeller. As illustrated in Figure 5,

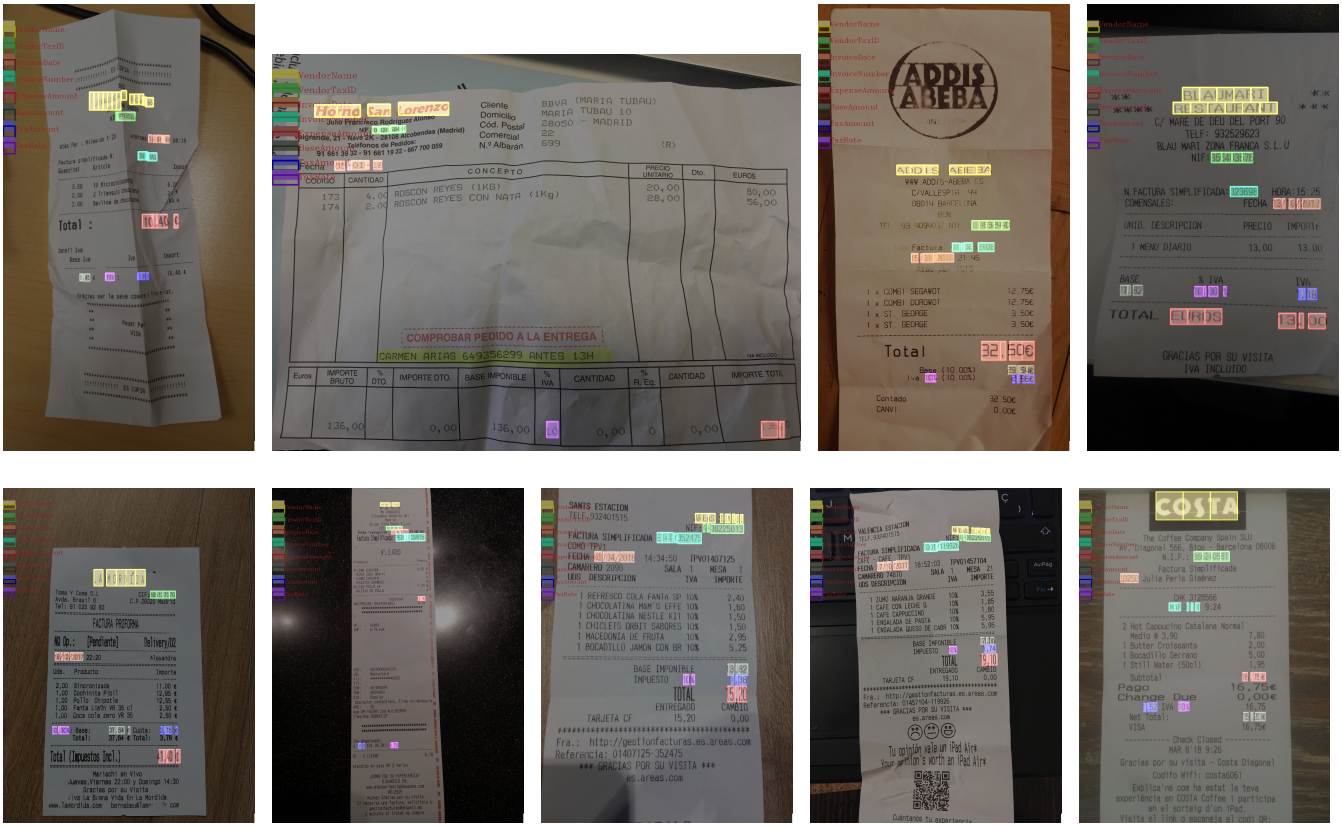

Figure 3. Example of inference results of CUITE. Color legend in the top-left corner indicates the key information classes. Each color indicates a key information class, where filled rectangles are the ground truths while the boundary-only rectangles are the inference results. The result is perfectly correct as if the filled rectangles overlap with the boundary-only rectangles.

4.3. Ablation Studies

Although we have demonstrated extremely strong empirical results, the results presented are achieved by combination of each aspect of the CUTIE framework. In this section, we perform ablation studies over a number of facets of CUTIE in order to better understand their relative importance.

4.3.1 Effect of Grid Augmentation on Understanding Spatial Information

One of our core claim is that the high performance CUTIE is achieved by jointly analysis of the semantical and spatial information with the proposed framework. The highly effective spatial analysis ability is enabled by the CUTIE framework in contrast with the previous NER based methods. The grid augmentation process further enhances this ability. To give evidence to this claim, we evaluate CUTIE with or without the grid augmentation process. Results are presented in Table 3. We can see that adding grid augment process in CUTIE can significantly improve the performance on meal entertainment receipt and taxi receipt. These results demonstrate that CUTIE can greatly benefit from enhancing data diversity in spatial distribution. For that reason, further analysis about grid augmentation techniques will surely further enhance CUTIE's performance, *e.g.*, randomly moving certain texts upward, downward, leftward, or rightward by several pixels during the grid positional mapping process.

Table 3. Performance evaluation of CUTIE with or without the grid augmentation process. (AP/softAP)

	MEnter	Taxi	Hotel
w augmentation	/	/	
w/o augmentation	/	/	

4.3.2 Impact of Semantical Information Capacity

Next, we evaluate the impact of semantical information of CUTIE by comparing evaluation with different word embedding sizes, although it is intuitively clear that larger em-

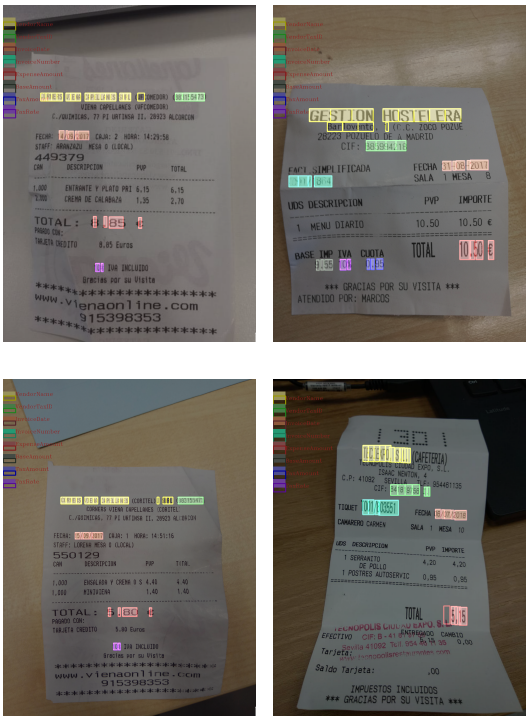

Figure 4. Example of prediction results of CUITE with false positive inference results. Color legend in the top-left corner indicates the key information classes. Each color indicates a key information class, where filled rectangles are the ground truths while the boundary-only rectangles are the inference results. The false positive are results with the boundary-only rectangles not overlap with the filled rectangles.

Table 4. Performance evaluation of CUTIE with different embedding size. (AP/softAP)

	64	128	256
	/	/	
	/	/	

bedding size provides more semantic capacity. As reported in Table 4, CUTIE performs worse with smaller embedding size. The result meets our expection that *details*. To further enhance the model performance, it might help to randomly mask out certain texts during the grid positional mapping process, such that the model learns inter-text relationship better with / without all words present in the grid and improves in generality.

5. Discussion

Automatically extracting interested words / information from the scanned document images is of great interest to various services and applications. The performance gain is

Figure 5. Example of prediction results of CUITE with false positive inference results. Color legend in the top-left corner indicates the key information classes. Each color indicates a key information class, where filled rectangles are the ground truths while the boundary-only rectangles are the inference results.

mainly achieved by exploring three key factors: the spatial relationship among texts, the semantic information of texts, and the grid positional mapping mechanism.

References

[1] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[2] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[3] L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[4] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[6] GithubBERT. https://github.com/google-research/bert.

[7] R. B. Palm, O. Winther, and F. Laws. Cloudscan - A configuration-free invoice analysis system using recurrent neural networks. *CoRR*, abs/1708.07403, 2017.

[8] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation, 2019.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

6. APPENDIX