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 method in 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 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 network and propose a faster and portable structure. We demonstrate the effectiveness of the proposed method on a dataset with up to 3000 labelled receipts, without any post-processing, achieving state-of-the-art performance that is higher than BERT but with only 1/10 parameters and much higher than the 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 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, which introduces great difficulties. The order or 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.

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 the formatted input of the entire network, which is arranged in close correspondance with their real spatial relationship of the original scanned document image. Furthermore, the rich semantic information is encoded from the gridded texts in the very beginning of the convolutional network.

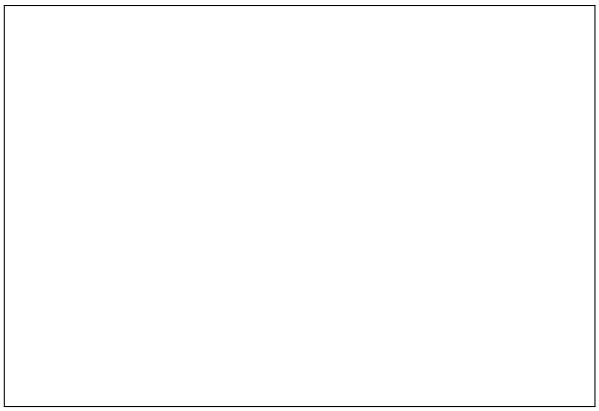

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

2. Related Work

3. Methods

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

3.1. Grid Positional Mapping

To generate input grid data for the convolutional 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 here will benefit the following 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 [5].

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 network.

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

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

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. 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 (texts) 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.

In fact, several methods have been proposed in the semantic segmentation field to capture multi-scale context 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 which is detrimental to the sparse grid. Moreover, the encoding and decoding process applies shape restricts to the grid shape augment process.

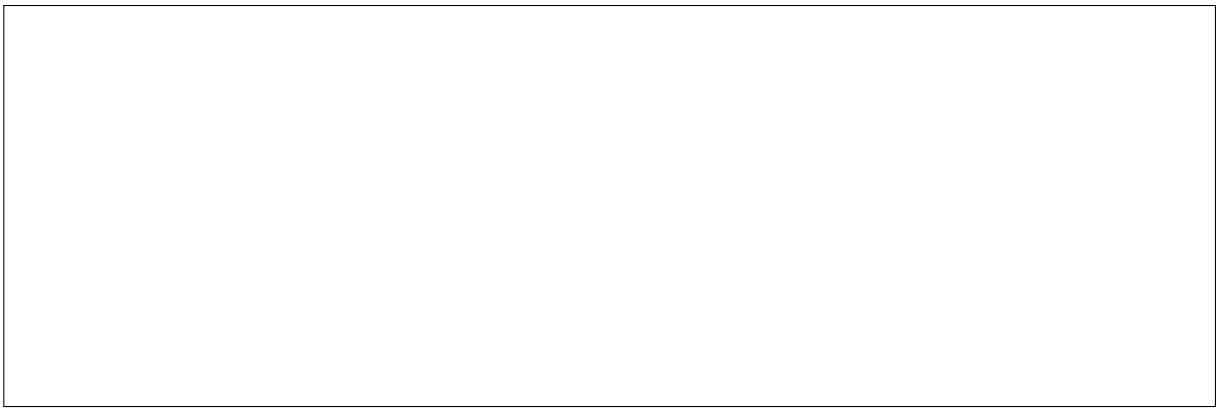

Figure 2. Example of a short caption, which should be centered.

Instead, the field of view of filters can also be effectively enlarged by combining multi-resolution features [6] 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 CUTIE-A to employ convolutional multi-resolution fusion network without losing high-resolution features and CUTIE-B network with atrous convolution.

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 [6], a high-resolution network without striding is employed as the backbone major 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.

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 $r = 1$.

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

Table 1. Performance comparison on 3 types of documents

Method	#Params	Meals	Taxi	hotel
CloudScan	-			
BERT	108M	79		
CUTIE-A	66M	82.5		
CUTIE-B	16M			

4. Experiments

The proposed method is evaluated on a dataset with with 3 types of documents which contain 9 key information classes and 1 don't care class. The dataset contains 4000 per-token annotated documents. We generate the text and bounding box features with Google's OCR API. The performance is measured in terms of per-token accuracy across the 10 classes.

We use a learning rate of 1e-3 for training with Adam optimizer with step decay learning strategy, and 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 each component. The embedding size is 128, target augmentation shape is 64 for both row and column.

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 mainly achieved by exploring three key factors: the spatial relationship among texts, the semantic information of texts, and the text gridding 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] B. Github. https://github.com/google-research/bert.

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