%%%%%%%%% TITLE

\title{CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor}

%%%%%%%%% ABSTRACT

\begin{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 $6,980$ labelled receipts, without any pre-training or post-processing, achieving state of the art performance that is much higher than BERT but with only $1/10$ parameters and without requiring the $3,300$M word dataset for pre-training.

\end{abstract}

%%%%%%%%% BODY TEXT

\section{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. \ref{fig:receipts}.

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. \ref{fig:cutie}. 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 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 \cite{bert,transformer}.