README.md

# CUTIE

TensorFlow implementation of the paper "CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor." Xiaohui Zhao ArXiv 2019

## Overview

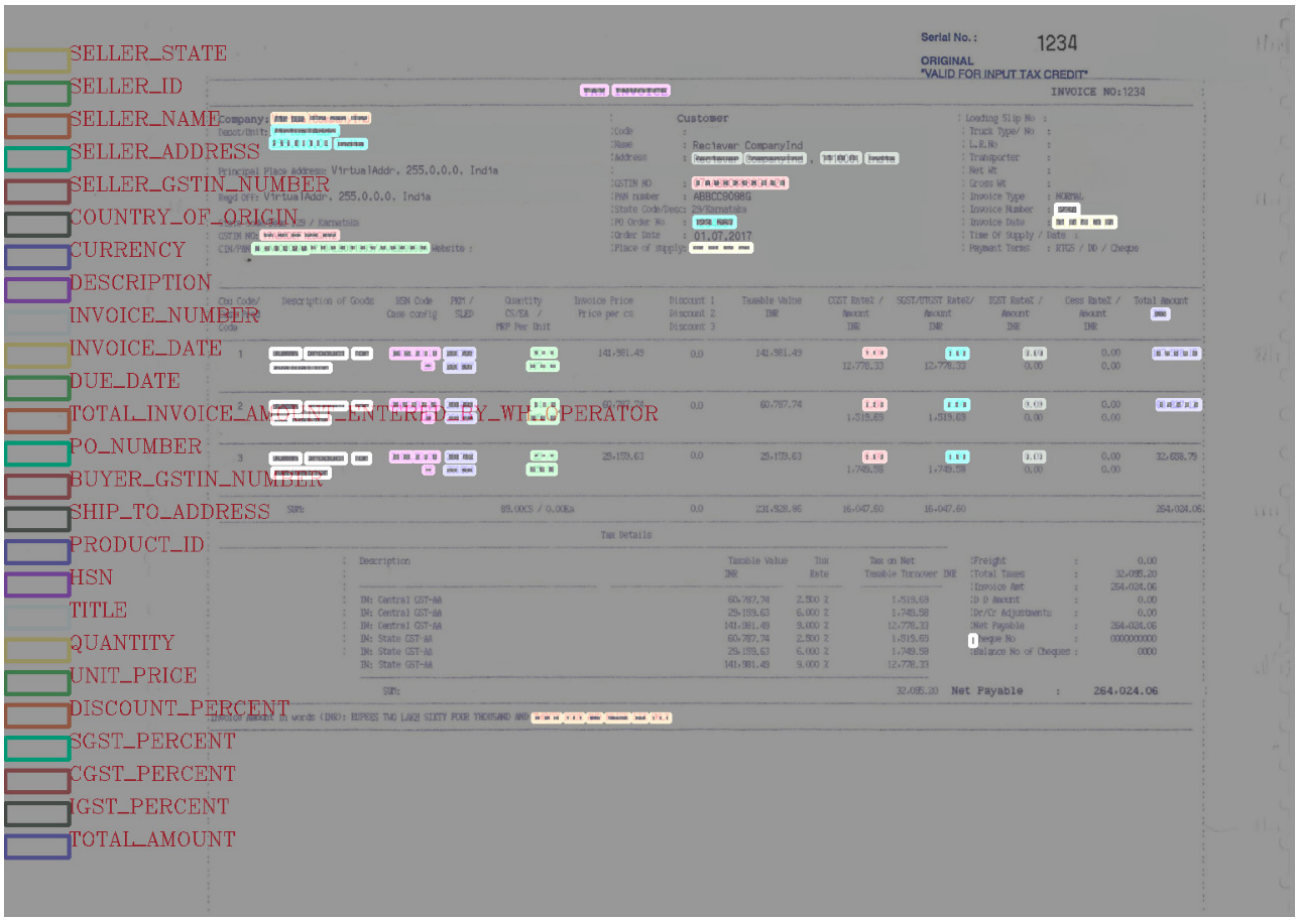
**CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor**

This paper proposes a learning-based key information extraction method with limited requirement of human resources. It combines the information from both semantic meaning and spatial distribution of texts in documents. Their proposed model, applies convolutional neural networks on gridded texts where texts are embedded as features with semantical connotations.

The proposed model, tackles the key information extraction problem by

- First creating gridded texts with the proposed **grid positional mapping method**. To generate the 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. The texts are mapped 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.
- Then the CUTIE model is applied on the gridded texts. 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.

Source: Nanonets




## Installation & Usage

```
pip install -r requirements.txt
```

1. Run `clovaai_api.py` for ocr on Train image dataset.
2. Using `textbox_generation.py` convert ocr json file to model compatible dataset.
3. Add remaining invoices field using `add_remianing.py` .
4. Open `dataset_creater.html` in browser to annotate the invoice fields.
5. Creat new vocab for your dataset using `create_vocab.py` .
6. Generate your own dictionary with main_build_dict.py / main_data_tokenizer.py
7. Train your model with main_train_json.py

CUTIE achieves best performance with rows/cols well configured. For more insights, refer to statistics in the file (others/TrainingStatistic.xlsx).

## Results

Result evaluated on 4,484 receipt documents, including taxi receipts, meals entertainment receipts, and hotel receipts, with 9 different key information classes. (AP / softAP)

| Method | #Params | Taxi | Hotel |
|---|---|---|---|
| CloudScan | - | 82.0 / - | 60.0 / - |
| BERT | 110M | 88.1 / - | 71.7 / - |
| CUTIE | **14M** | **94.0** / **97.3** | **74.6** / **87.0** |