Intelligent document information extraction based on convolutional neural network (CNN), natural language processing (NLP) and graph neural network (GNN)

Université de Haute Alsace

Mouad HAMRI\textsuperscript{1,2}  Maxime Devanne\textsuperscript{1}  Jonathan Weber\textsuperscript{1}  Michel Hassenforder\textsuperscript{1}

1. IRIMAS, Université de Haute-Alsace, Mulhouse, France
2. Syentys, France

August 21, 2022
Agenda

1. Introduction

2. State of the art & Related works

3. Enhancing GNN feature modeling for document information extraction using transformers

4. Conclusion and future work
Table of Contents

1. Introduction
2. State of the art & Related works
3. Enhancing GNN feature modeling for document information extraction using transformers
4. Conclusion and future work
Problem statement

- Form-like documents as invoices are massively used in day-to-day business workflow.
- Problems to solve:
  - Extract key-value information such as date, supplier name and amount.
  - Detect tables and extract lines key-values (e.g. invoice lines: description, quantity, unit amount, ...).
Figure: Invoice information parsing. Figure reproduced from: ¹

¹https://cloud.google.com/document-ai/docs/invoice-parser
Problem statement

Regular expression, heuristics and rules patterns
- Require big effort and time to configure the models.
- Bad performance for unseen templates.

Deep learning models
- Better results with less effort.
- Generalise better for unseen templates
Table of Contents

1 Introduction

2 State of the art & Related works

3 Enhancing GNN feature modeling for document information extraction using transformers

4 Conclusion and future work
Figure: CNN architecture proposed by LeCun et al. Figure reproduced from: \textsuperscript{2}

\textsuperscript{2}Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE 86, no. 11 (1998)
Figure: RNN architecture.  

M. VENKATACHALAM *Recurrent Neural Networks – Remembering what’s important.* Figure reproduced from: URL https://gotensor.com/2019/02/28/recurrent-neural-networks-remembering-whats-important/
The Transformer model architecture.

Graph neural networks (GNN) - relation processing

There are three families of tasks most GNN models are trained on:

- **Graph classification**: e.g. molecule property prediction.
- **Node classification**: e.g. classify paper topic in a citation graph.
- **Link prediction**: e.g. predict if two users in a social network have a link.
Graph neural networks (GNN)

GNN Convolution

Two families of convolution models:

- Spectral Convolution: GNN convolution operator based on the graph Fourier transform\(^a\).
- Spatial Convolution: use the graph topology to learn node representation using its neighborhood.

Figure: Graph neural network node aggregation. The figure is reproduced from:  

Related works (deep learning)

Many works $^a$ $^b$ $^c$ used:

- Text embedding.
- Visual embedding.
- Spatial features (e.g. bounding boxes coordinates).

---


State of the art & Related works

LayoutLMv2 architecture.

State of the art & Related works

Related works (GNN)

The usage of a GNN model to extract document information has been already adopted in many works:

- \(^a\): based on Chebyshev graph convolutional neural networks \(^b\).
- \(^c\): based on the Graph Attention Network (GAT) \(^d\)


# Table of Contents

1. Introduction

2. State of the art & Related works

3. Enhancing GNN feature modeling for document information extraction using transformers

4. Conclusion and future work
GNN document representation

GNN Vs CNN & NLP

- The representation of the document using a graph is a better approach as the graph will preserve the document structure.
- The graph representation is much more efficient compared to pixel-grid representation (CNN) as the storage of the graph requires less memory comparing to the document image pixels.
Graph construction

Features assignment

We assign a feature vector to each extracted text / bounding box from the previous step. The features are assigned as follow:

- **Spacial features**: a vector of 3 floats composed of the normalised coordinates of the box center: $x_{center}/w$ and $y_{center}/h$ ($w$ and $h$ are the width and height of the image respectively) and the normalized box line number: $l_{box}/L$ ($l_{box}$ is the box line number and $L$ is document number of lines).

- **Text features**: a vector of 6 floats representing the number of lower, upper, special, alphanumeric, numeric and space characters in the box text. The vector is normalized by dividing each element by the max occurrence in the document.
Figure: Example of graph neighbours selection and receipt graph construction
Model architecture

Figure: Model architecture using GNN transformer layers

\[ \frac{\exp(x_i)}{\sum_j \exp(x_j)} \]

\( \text{Softmax} \)

\( \text{Relu} \)

4 transformers layers (\# heads=4)

Input graph

Classified nodes

Grid search - Hyper-parameters

- 4 layers of "graph transformers"
- 4 heads
- Hidden size of 16
- ADAM optimizer (learning rate of 0.001 / weight decay of $5 \times 10^{-4}$)
- Dropout of 0.1
Model pipeline

Figure: Pipeline of our model
Figure: Example of SROIE dataset receipt image, bounding boxes and fields values (626 documents for training and 347 for testing)  

6https://rrc.cvc.uab.es/?ch=13
Variants

Figure: Features variants: LM777 is the combination of LM768 and ST9 while BERT777 is the combination of BERT768 and ST9
## Results

**Figure:** Precision, Recall and F1 scores result on SROIE dataset for LM777 model

<table>
<thead>
<tr>
<th>Seed</th>
<th>K1</th>
<th>K2</th>
<th>K3</th>
<th>K4</th>
<th>K5</th>
<th>MEAN</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>0</td>
<td>0.954</td>
<td>0.919</td>
<td>0.936</td>
<td>0.933</td>
<td>0.893</td>
<td>0.913</td>
<td>0.923</td>
</tr>
<tr>
<td>1</td>
<td>0.935</td>
<td>0.887</td>
<td>0.910</td>
<td>0.937</td>
<td>0.892</td>
<td>0.914</td>
<td>0.932</td>
</tr>
<tr>
<td>2</td>
<td>0.950</td>
<td>0.900</td>
<td>0.924</td>
<td>0.943</td>
<td>0.893</td>
<td>0.917</td>
<td>0.926</td>
</tr>
<tr>
<td>3</td>
<td>0.935</td>
<td>0.886</td>
<td>0.910</td>
<td>0.935</td>
<td>0.897</td>
<td>0.916</td>
<td>0.932</td>
</tr>
<tr>
<td>4</td>
<td>0.930</td>
<td>0.900</td>
<td>0.915</td>
<td>0.929</td>
<td>0.883</td>
<td>0.906</td>
<td>0.938</td>
</tr>
</tbody>
</table>

OA  0.937 | 0.891 | 0.9132 | 0.009 | 0.009 | 0.0084 | 0.936
Table: Precision, Recall and F1 scores result on SROIE dataset of our model compared to the other variants. The bold values represent the higher results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>$\mu$ (F1)</th>
<th>$\sigma$ (F1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM777</td>
<td>0.9542</td>
<td>0.9191</td>
<td>0.9363</td>
<td>0.9132</td>
<td>0.0084</td>
</tr>
<tr>
<td>BERT777</td>
<td>0.9467</td>
<td>0.9104</td>
<td>0.9282</td>
<td>0.9111</td>
<td>0.0092</td>
</tr>
<tr>
<td>LM768</td>
<td>0.9533</td>
<td>0.9003</td>
<td>0.9260</td>
<td>0.9106</td>
<td>0.0072</td>
</tr>
<tr>
<td>BERT768</td>
<td>0.9505</td>
<td>0.9025</td>
<td>0.9259</td>
<td>0.9113</td>
<td>0.0096</td>
</tr>
<tr>
<td>ST9</td>
<td>0.8601</td>
<td>0.7818</td>
<td>0.8195</td>
<td>0.8009</td>
<td>0.0114</td>
</tr>
</tbody>
</table>
Results

- F1 score of 0.9363 on SROIE dataset.
- Best result achieved using a purely GNN model to our knowledge.
Results

SROIE Top 1 (StrucTexT)
- F1 score: 98.70%
- Trained on 900k documents.
- 107M parameters.

Our model
- F1 score: 93.63%
- Trained on 500 receipts (0.055% in comparison with SROIE Top 1).
- 53.6K parameters (training phase) (0.050% in comparison with SROIE Top 1).
1. Introduction

2. State of the art & Related works

3. Enhancing GNN feature modeling for document information extraction using transformers

4. Conclusion and future work
Conclusion

- Graph neural network model to label Form-like documents fields.
- Optimized method to construct the document graph while having smaller graphs with rich node embedding based on LayoutLMv2 representation combined with spacial and textual features.
- Multi graph transformer layers to force the model to focus on the most relevant neighbours.
- F1 score of 0.9363 on SROIE dataset (best result achieved using a purely GNN model to our knowledge).
Future work

- Conceive new graph convolutional layers suitable for document information extraction problem while fine-tuning the features and graph architectures.
- Use larger datasets.
Thank you!

Contact
mouad.hamri@uha.fr