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[1,2] Maxime Devanne[1] Jonathan Weber[1] Michel Hassenforder[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 extration 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 extration 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, ...).

Problem statement

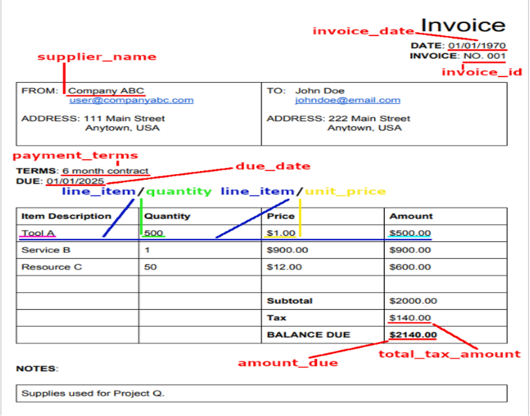


Figure: Invoice information parsing. Figure reproduced from: [1]

[1] https://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 extration using transformers
4. Conclusion and future work

Convolutional neural networks (CNN) - image processing

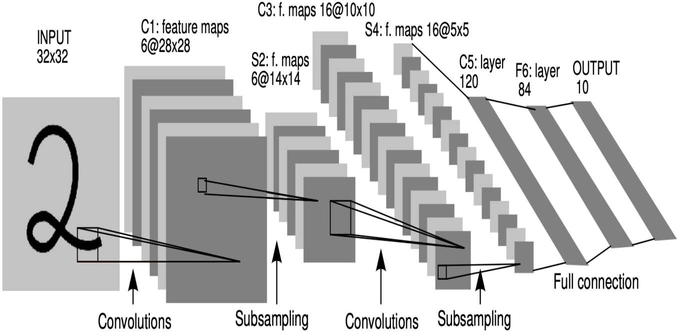


Figure: CNN architecture proposed by LeCun et al. Figure reproduced from: [2]

[2] Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE 86, no. 11 (1998)

RNN - Text processing

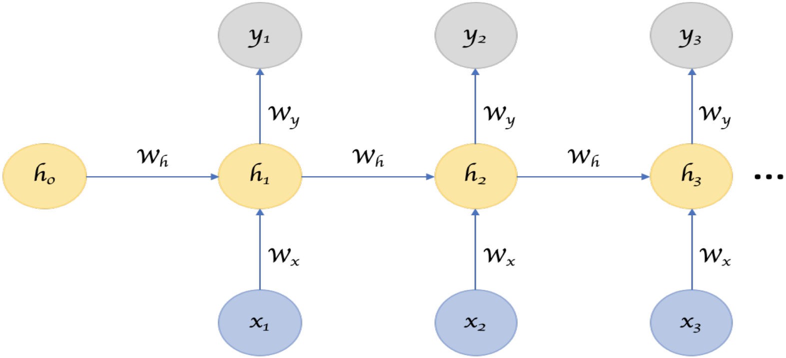
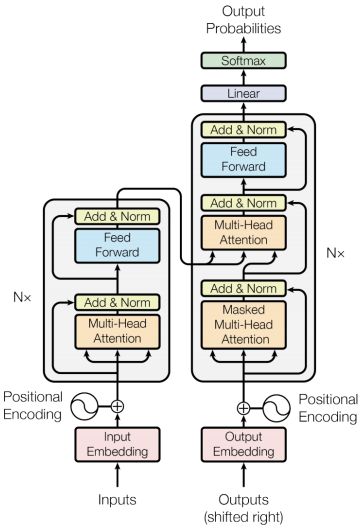


Figure: RNN architecture. [3]

[3] 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/

Transformers

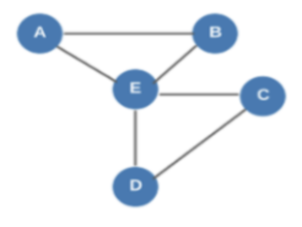


The Transformer model architecture.

Figure reproduced from: Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, et Illia Polosukhin. *Attention Is All You Need* 2017 arXiv:1706.03762 [cs] http://arxiv.org/abs/1706.03762

Graph neural networks (GNN) - relation processing

Graph



GNN tasks

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.

[a] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega and P. Vandergheynst *The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains* IEEE Signal Processing Magazine, vol. 30, no. 3, pp. 83-98, May 2013, doi: 10.1109/MSP.2012.2235192

Spatial Convolution

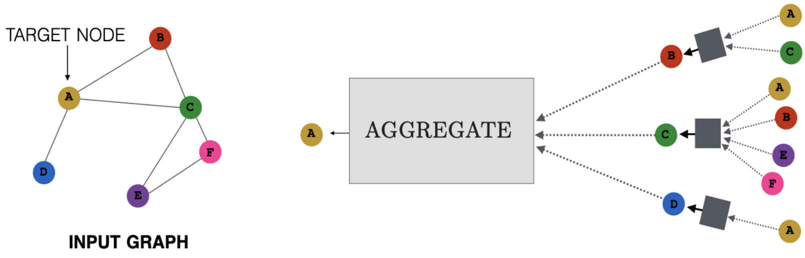


Figure: Graph neural network node aggregation. The figure is reproduced from: [4]

[4] Hamilton, William L. *Graph Representation Learning* Synthesis Lectures on Artificial Intelligence and Machine Learning. volume 14, number 3,1-159, Morgan and Claypool.

State of the art & Related works

Related works (deep learning)

Many works [a] [b] [c] used:

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

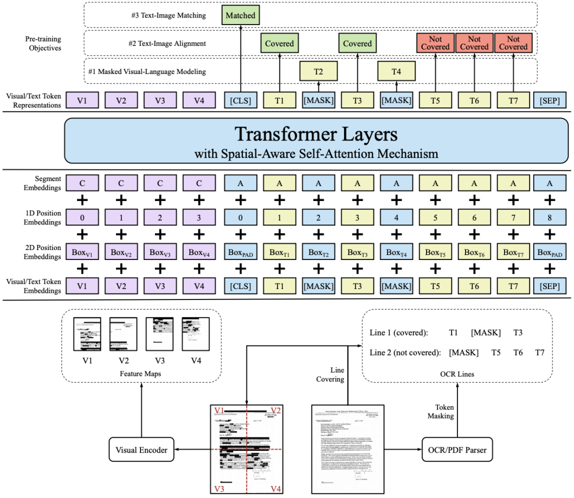
---

[a] R. B. Palm, F. Laws, et O. Winther, *Attend, Copy, Parse End-to-end Information Extraction from Documents* in 2019 International Conference on Document Analysis and Recognition (ICDAR), sept. 2019, p. 329-336. doi: 10.1109/ICDAR.2019.00060.

[b] S. Patel et D. Bhatt, *Abstractive Information Extraction from Scanned Invoices (AIESI) using End-to-end Sequential Approach* arXiv:2009.05728 [cs], sept. 2020

[c] Timo I. Denk and Christian Reisswig *BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding* Workshop on Document Intelligence at NeurIPS 2019, 2019.

State of the art & Related works



LayoutLMv2 architecture.

The figure is reproduced from: Xu, Yang, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, et al. *LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding* arXiv:2012.14740 [cs], http://arxiv.org/abs/2012.14740

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]

[a] Lohani,D.,Belaïd,A.,Belaïd,Y. *An invoice reading system using a graph convolutional network* Carneiro, G., You, S. (eds.) ACCV 2018. LNCS, vol. 11367, pp. 144–158. Springer, Cham (2019)

[b] Michael Defferrard, Xavier Bresson, and Pierre Vandergheynst *Convolutional neural networks on graphs with fast localized spectral filtering*. In Advances in Neural Information Processing Systems, pp. 3844–3852, 2016

[c] Belhadj D., Belaïd Y., Belaïd A. *(2021) Consideration of the Word's Neighborhood in GATs for Information Extraction in Semi-structured Documents* In: Lladós J., Lopresti D., Uchida S. (eds) Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science, vol 12822. Springer, Cham. https://doi.org/10.1007/978-3-030-86331-9_55

[d] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio *Graph attention networks*. Proc. ICLR, pp. 1-12, 2017.

Table of Contents

1 Introduction

2 State of the art & Related works

3 Enhancing GNN feature modeling for document information extration 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.

Graph construction

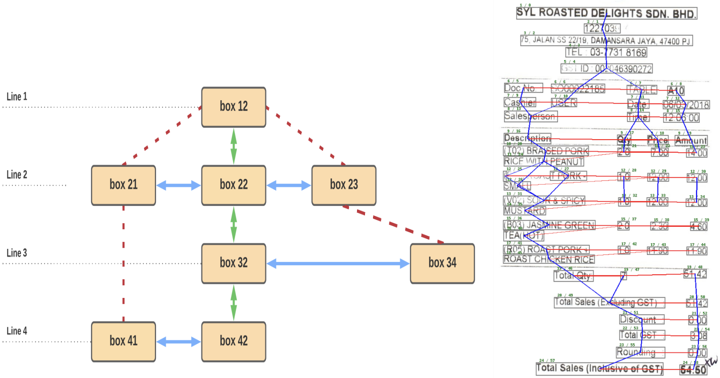


Figure: Example of graph neighbours selection and receipt graph construction

Model architecture

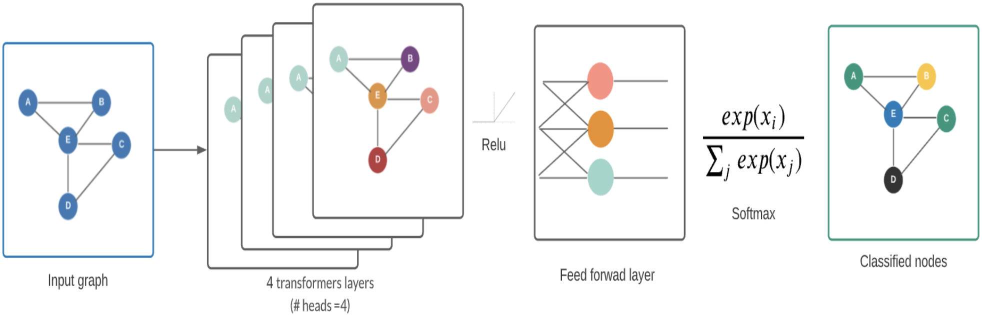


Figure: Model architecture using GNN transformer layers[5]

[5] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, et Y. Sun, *Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification* arXiv:2009.03509 [cs, stat], mai 2021, http://arxiv.org/abs/2009.03509

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 5×10^{-4})
- Dropout of 0,1

Model pipeline

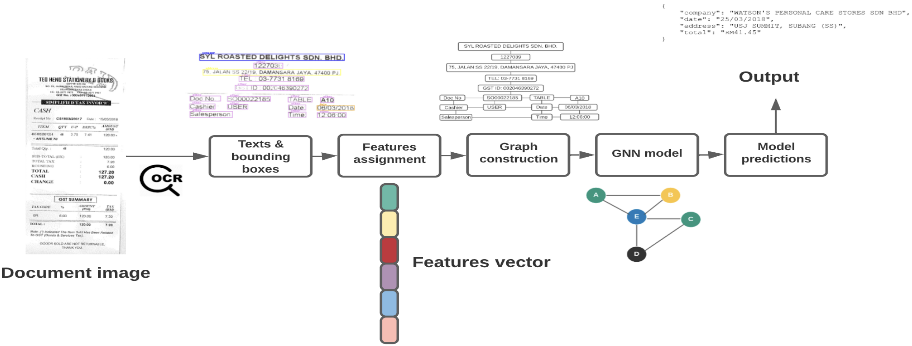


Figure: Pipeline of our model

Dataset

Figure: Example of SROIE dataset receipt image, bounding boxes and fields values (626 documents for training and 347 for testing) [7]

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

Variants

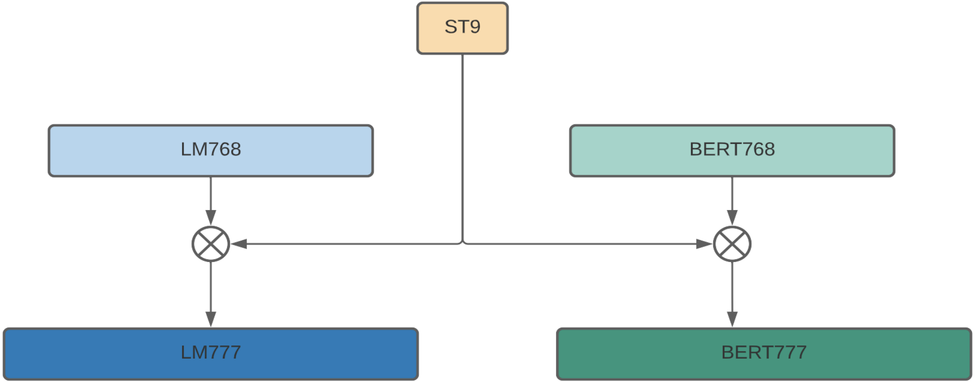


Figure: Features variants: LM777 is the combination of LM768 and ST9 while BERT777 is the combination of BERT768 and ST9

Results

	K1			K2			K3			K4			K5			MEAN			STD			
Seed	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	**F1**	P	R	**F1**	Max F1
0	**0,954**	**0,919**	**0,936**	0,933	0,893	0,913	0,923	0,889	0,906	0,941	0,885	0,912	0,929	0,878	0,903	0,936	0,893	**0,914**	0,012	0,016	**0,013**	**0,936**
1	0,935	0,887	0,910	0,937	0,892	0,914	0,932	0,893	0,912	0,942	0,891	0,916	0,936	0,894	0,914	0,936	0,891	**0,913**	0,004	0,003	**0,002**	**0,916**
2	0,950	0,900	0,924	0,943	0,893	0,917	0,926	0,881	0,903	0,943	0,889	0,915	0,948	0,886	0,916	0,942	0,890	**0,915**	0,010	0,007	**0,008**	**0,924**
3	0,935	0,886	0,910	0,935	0,897	0,916	0,932	0,880	0,905	0,916	0,872	0,894	0,946	0,896	0,920	0,933	0,886	**0,909**	0,011	0,011	**0,010**	**0,920**
4	0,930	0,900	0,915	0,929	0,883	0,906	0,938	0,889	0,913	0,947	0,912	0,929	0,939	0,888	0,913	0,937	0,894	**0,915**	0,007	0,011	**0,009**	**0,929**
															OA	0,937	0,891	**0,9132**	0,009	0,009	**0,0084**	**0,936**

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

Results

Results

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

Model	Precision	Recall	F1	μ (F1)	σ (F1)
LM777	**0,9542**	**0,9191**	**0,9363**	0,9132	0,0084
BERT777	0,9467	0,9104	0,9282	0,9111	0,0092
LM768	0,9533	0,9003	0,9260	0,9106	0,0072
BERT768	0,9505	0,9025	0,9259	0,9113	0,0096
ST9	0,8601	0,7818	0,8195	0,8009	0,0114

Results

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).

Table of Contents

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

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).

Conclusion and future work

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