Doc2Graph: a Task Agnostic Document Understanding Framework based on Graph Neural Networks

Andrea Gemelli[1], Sanket Biswas[2], Enrico Civitelli[1], Josep Lladós[2], Simone Marinai[1]

[1]*Dipartimento di Ingegneria dell'informazione (DINFO), Università degli studi di Firenze, Italy*

[2]*Computer Vision Center & Computer Science Department, Universitat Autònoma de Barcelona, Spain*

Overview

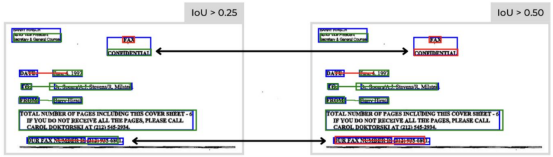
Document Understanding (DU) comprises several challenging tasks, usually achieved by several and complex models.
Graph Neural Networks (GNNs) have been proposed to solve such task but lacking of a general representation of documents as graphs and relying on heuristics.

Our **Doc2Graph** proposes:

- A task-agnostic GNN-based DU framework, evaluated on two challenging benchmarks (forms and invoices) for three significant tasks;
- propose a general graph representation module for documents, that do not rely on heuristics to build pairwise relationships between words or entities;
- A novel GNN architectural pipeline with node and edge aggregation functions suited for documents, that exploits the relative positioning of document objects through polar coordinates

Tokens (nodes) granularity

Tokens, graph nodes, can be either entities or words.
Results of Entity Detection using YOLOv5-small[C]



	Metrics (↑)			*% Drop Rate* (↓)	
IoU	**Precision**	**Recall**	**F_1**	**Entity**	**Link**
0.25	0.8728	0.8712	0.8720	12.72	16.63
0.50	0.8132	0.8109	0.8121	18.67	25.93

We apply an OCR to consider words instead.

References

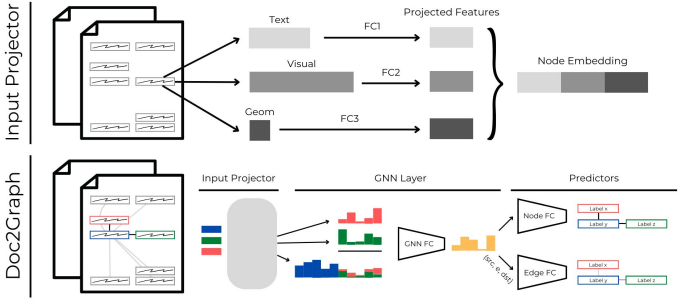
[A] *Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 1-6. IEEE (2019)*

[B] *Riba, P., Dutta, A., Goldmann, L., Fornes, A., Ramos, O., Lladós, J.: Table detection in invoice documents by graph neural networks. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 122–127. IEEE (2019)*

[C] *Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Fang, J., imyhxy, Michael, K.: ultralytics/yolov5: v6. 1-tensorrt, tensorflow edge tpu and openvino export and inference. Zenodo, Feb 22 (2022)*

UNIVERSITÀ DEGLI STUDI FIRENZE · CVC Centre de Visió per Computador · ECCV TEL AVIV 2022 · Code

Model and Pipeline

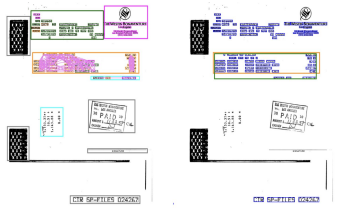


We tried different variations of our model, using different features:

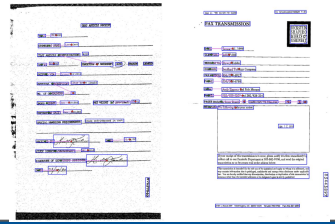
Features					*F_1 per classes* (↑)			
Geometric	**Text**	**Visual**	**EP Inner dim**	**IP FC dim**	**None**	**Key-Value**	**AUC-PR** (↑)	**# Params** $\times 10^6$ (↓)
✓	✗	✗	20	100	0.9587	0.1507	0.6301	0.025
✗	✓	✗	20	100	0.9893	0.1981	0.5605	0.054
✓	✓	✗	20	100	0.9941	0.4305	0.7002	0.120
✓	✓	✗	300	300	0.9961	0.5606	0.7733	1.18
✓	✓	✓	300	300	**0.9964**	**0.5895**	**0.7903**	2.68

Qualitative results

Layout Analysis and Table Detection



Entity Labeling and Linking



Quantitative Results

Results on FUNSD dataset[A]

		F_1 (↑)		
Method	**GNN**	**Semantic Entity Labeling**	**Entity Linking**	**# Params** $\times 10^6$ (↓)
BROS [13]	✗	**0.8121**	**0.6696**	138
LayoutLM [33,13]	✗	0.7895	0.4281	343
FUNSD [15]	✓	0.5700	0.0400	-
Carbonell et al. [6]	✓	0.6400	0.3900	201
FUDGE w/o GCN [8]	✗	0.6507	0.5241	12
FUDGE [8]	✓	0.6652	0.5662	17
Doc2Graph + YOLO	✓	0.6581 ± 0.006	0.3882 ± 0.028	13.5
Doc2Graph + GT	✓	0.8225 ± 0.005	0.5336 ± 0.036	**6.2**

Results on RVL-CDIP dataset[B], for Layout Analysis and Table Detection

	Accuracy (↑)	
Method	**Max**	**Mean**
Riba et al. [25]	62.30	-
Doc2Graph + OCR	**69.80**	**67.80** ± 1.1

		Metrics (↑)		
Method	**Threshold**	**Precision**	**Recall**	**F_1**
Riba et al. [25]	0.1	0.2520	**0.3960**	0.3080
Riba et al. [25]	0.5	0.1520	0.3650	0.2150
Doc2Graph + OCR	0.5	**0.3786** ± 0.07	0.3723 ± 0.07	**0.3754** ± 0.07

Conclusions and Future Works

- We have presented a task-agnostic document understanding framework based on a Graph Neural Network;
- We have proposed a general representation of documents as graphs, exploiting fully connectivity between document objects;
- Our light-weight model can achieve promising results over three tasks and two challenging datasets;
- For the future, we will consider extend the range of the possible application of the library, including multi-lingual tasks.

Acknowledgments

This work has been partially supported by the the Spanish projects MIRANDA RTI2018-095645-B-C21 and GRAIL PID2021-126808OB-I00, the CERCA Program / Generalitat de Catalunya, the FCT-19-15244, and PhD Scholarship from AGAUR (2021FIB-10010).

Contact: andrea.gemelli@unifi.it