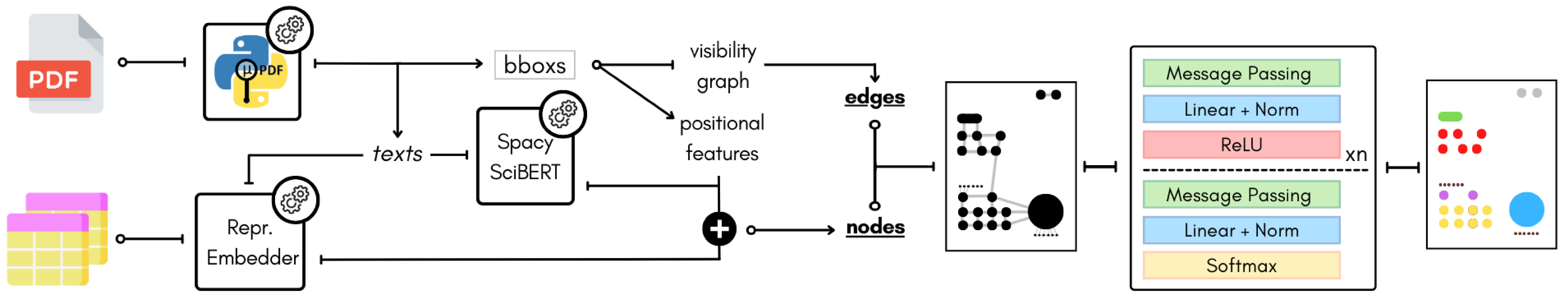
Graph Neural Networks and Representation Embedding for Table Extraction in PDF Documents

Andrea Gemelli*, Emanuele Vivoli* and Simone Marinai
Dipartimento d'Ingegneria dell'Informazione, AILab, Università degli studi di Firenze, Firenze
contacts: {andrea.gemelli, emanuele.vivoli}@unifi.it

Introduction

Problem Table detection, recognition and functional analysis are often performed separately, without a context.
Contributions We redefine the Table Extraction task [2] as a token classification one, exploiting Graph Neural Networks to perform node classification (detecting tables, table headers and other document objects) and edge classification (recognizing table structure). We directly make use of PDF documents and enrich node vectors with novel representation embeddings.

Method



Network The Message Passing in use is similar to [1] with mean aggregator function. The contribution of neighbours is scaled given their distance from source node.
Graphs are generated from PDF files in three steps:

1. tokens (nodes) in PDFs are extracted using PyMuPDF, a Python binding for MuPDF;
2. each node is connected using a visibility graph;
3. nodes have geometrical, representation and text feature embeddings; edges have distances between connected nodes.

Representation embedding Each node obtain its representation embedding: (i) mapping the content into a standardized representation (e.g. "*Precision-Recall*" → "*w-w*"; "12.5" → "*x.x*"); (ii) embedding the representation in a dense vector.

References

[1] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *NIPS17*, 30, 2017.

[2] B. Smock, R. Pesala, and R. Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proc. CVPR22, pp.4634-4642*, 2022.

[3] X. Zhong, J. Tang, and A. J. Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019.

Results

Features	Metrics	Methods		
		Base	*Padding*	*Scaled*
bbox	accuracy	0.873	0.841	0.866
	cell F1	0.798	0.765	0.799
	cell-h F1	0.659	0.651	0.642
bbox + repr	accuracy	0.876	**0.875**	0.873
	cell F1	0.821	0.819	0.816
	cell-h F1	0.653	0.649	0.648
bbox + Spacy	accuracy	0.859	0.847	0.868
	cell F1	0.767	0.773	0.781
	cell-h F1	0.685	0.675	0.660
bbox + repr + Spacy	accuracy	0.865	0.860	0.809
	cell F1	0.780	0.776	0.811
	cell-h F1	**0.689**	0.675	0.644
bbox + SciBERT	accuracy	**0.882**	0.843	**0.879**
	cell F1	0.838	0.816	**0.846**
	cell-h F1	0.688	**0.699**	**0.686**
bbox + repr + SciBERT	accuracy	0.709	0.787	0.870
	cell F1	**0.855**	**0.832**	0.777
	cell-h F1	0.668	0.636	0.671

Data custom subset of PubLayNet [3] and PubTables-1M [2] datasets.

Conclusions

(i) **redefined the problem of table extraction**: as a token classification task, exploiting a GNN and through a pipeline to represent PDFs as graphs.

(ii) **novel representation embeddings**: we have shown them to be effective to discriminate table elements from other classes.

(iii) **new collection of annotations and data** is proposed to address the table extraction task, adding two more classes 'caption' and 'page-info'.