Progress Report Ph.D.
Curriculum Smart Computing

Andrea Gemelli

September 2022

Contents

UNIVERSITÀ
DEGLI STUDI
FIRENZE

1 First Year - 2020/2021

1.1 Achievements

In the first year of my PhD I initially continued my master thesis work whose main objective has been to perform administrative document understanding exploiting geometric deep learning methods. In particular, the main goal has been to build a pipeline capable to extract the graph document structures in order to find key information inside invoices, such as sender, receiver and total amount to be paid. Graphs proved themself to be well suited for that, helping to bring also structural features in the deep learning phase. Starting from the work described in [14], we proposed and compared different techniques in the graph construction pipeline and a novel graph convolutional architecture. We have been able to reproduce and obtain slightly better results on their provided subset of RVL-CDIP (518 annotated documents), mainly due to some structural information added to the graph edges.

One important limitation of the research in administrative document understanding is the lack of large publicly available datasets, with meaningful/useful labels and high quality documents. One possibility to continue working in that direction would be to explore new data augmentation or auto-labelling techniques, or involving methods that can deal with zero or few examples, e.g. zero/few shot learning. Therefore, in order to continue studying and using graphs in the document analysis domain, we extended the subject of study to the important field of table understanding in scientific papers. The reasons have been mainly two: in the domain of scientific papers large datasets such as PubLayNet [18] and DocBank [10] are available and, an in-depth study on objects such as tables does not exclude being able to reapply it again in the future to administrative documents.

The graphs previously created to represent an entire document have been reused to represent table structures. The first experiments conducted involved Detectron2 to automatically find cell objects inside tables: subsequently, the graph has been constructed connecting their bounding boxes using a visibility approach. In the nearly future I will try also to enrich nodes feature vectors with content information (with NLP-based techniques) and deep visual features: bringing different approaches together have already been reached great results [17].

In the first year I also extensively studied the state of the art of topics of interest for my research, also thanks to the conferences I attended (Sec. 1.3). I first deepened my knowledge about graph neural networks starting from the very first works [15] until the most recent discoveries [3, 4]. In the last three years graph have gathered an astonishing and increasing attention, allowing the publication of a growing number of papers in a wide range of different topics, along with new specialized libraries and benchmark datasets. Among others domain like chemistry, also researchers from the document analysis community brought graphs in their works [13, 14, 17]

In parallel, I also studied in depth the problem of table understanding, meaning both detection among other document objects and recognition of their structure in the meaning of cells, rows and columns. [8]. While the first task is nowadays a mostly solved task, the second one is still a challenging problem: simply applying an OCR is not enough to make them completely understandable and editable. Once the structure has been properly recognized, there are several consequent applications that can be performed, such as keep track of the state-of-the-art models among the literature exploiting results tables [9]. In the future, I would also like trying to apply my acquired knowledge in order to help visually impaired people access document contents in a smarter and easier way.

Most recently I also started to study the latest discoveries in the natural language processing domain, mainly connected to the document analysis one.

1.2 Plans and Future Work

As my main purpose for the next research year, I would like to continue and improve what I started during this one, focusing on tables and scientific papers. Among the steps I would like to take in the near future there is to improve and continue the work already done with tables:

- adding more relevant information to the nodes of table graph structures, also using NLP-based approaches;
- enriching table content finding in the paragraphs where the table is cited additional information exploiting in-paper references, e.g. information about datasets, models and metrics proposed;

Moreover, looking forward to the next two years, it would be interesting find a complete and summarised representation of an entire scientific paper in order to perform information extraction, topic classification and other related tasks.

1.3 Conferences and Summer Schools

I have attended the following courses, soft and complementary skills and conferences:

- NEURIPS 2020. 6-12 Dec 2020.
- ICPR 2020. 10-15 Jan 2021.
- SSPR 2020. 21-22 Jan 2021.
- ”Learning Symbolic Equations with Deep Learning”, ACM. 7 Jun 2021.
- ICDAR 2021. 5-10 Sep 2021.
- UCA Deep Learning School 2021, Université Côte D’Azur: Graph neural networks and neural-symbolic computation, (Marco Gori)

- 4th IAPR SSDA 2021 (on-site), Luleå University of Technology, 23-27 Aug: recent developments in document analysis.

2 Second Year - 2021/2022

2.1 Achievements

During my second year my research kept focusing on the application of Geometric Deep Learning to the domain of Document Analysis. I have mainly addressed the "plans and future work" outlined at the end of the first one (Sec. 1.2). We found these limitations interesting to be further explored:

1. scarcity of data for administrative documents;
2. usually table extraction task is tackled do not considering also other contextual information.
3. lack of information carried out throughout the graph structure beyond layout positioning, such as language and visual features;

To address the scarcity of data, we published two papers with novel data augmentation techniques. The first one, "Data augmentation on graphs for table type classification" [2], proposes data augmentation directly on the graph structure. We have shown that, applying a combination of node /edge removal and column/row inversion techniques, we were able to re-balance the dataset in our hand and increase the performances for the downstream task. On the contrary in "Automatic generation of scientific papers for data augmentation in document layout analysis" [12], we have exploited a LayoutTransformer to generate scientific paper pages, both single and double column layouts. Using the new generated data, helped improving AP scores for Document Layout Analysis over two small collection of papers, ICDAR 2019 and of the ICPR 2021 workshops, in particular for double column layouts.

Regarding the table extraction task, we found out that several work carry out it in different separate steps. In "Graph neural networks and representation embedding for table extraction in PDF documents" [6], we propose a geometric approach to tackle the Contextualized Table Extraction problem, addressing table extraction and document layout analysis at once. We also enriched nodes with novel "representation embeddings" and our ablation studies have shown that they are a good alternative to language models, such as SpaCy [5] or SciBERT [1], to distinguish tables from the rest of the document objects.

During my research stay at the Computer Vision Center (CVC) I deepened my knowledge about GNNs applied over documents related tasks, working to a new project with experts in this field. I developed a new library called "Doc2Graph", to create a task-agnostic pipeline overcoming the limitations that usually are applied during the extraction of graph structures from documents. We validate this novel proposal for four different task over two challenging benchmarks for Information Extraction and Document Understanding in our paper "Doc2Graph: a Task Agnostic Document Under- standing Framework based on Graph Neural Networks" [7]. We compared with other SOTA models, both graphs and transformers, obtaining good results with a small amount of trainable parameter; furthermore, classifying nodes and edges in a end-to-end manner, we brought interpretability over the graph structure (in particular for the Entity Linking task on FUNSD). Doc2Graph helped also to address the third point left as to be done from the first year: we used visual features applying and comparing different backbones, and we have proposed novel relative positioning features over edges. In the paper we extensively discuss their usage and usefulness through ablation studies.

Beyond the work carried on, the experiences I have made positively helped me to increase my knowledge and networking. As a visiting student to the CVC I could work with expert colleagues and improve my research work, also with other topics of interest of machine learning in general. Furthermore, being in presence at ICPR conference and S+SSPR workshop, gave me the opportunity to talk about my work over a larger audience of expertise all over the world.

2.2 Plans and Future Work

For my next and last year of Ph.D. i would like to further explore these points:

- extend the data augmentation works over business documents. A new dataset ”DocLayNet” [11] has been just released, and could be interesting to explore it to try the already developed techniques;

- try and propose novel ”anonymous properties” for business documents as an alternative to language models, given the interest of industry over privacy issues;

- Expanding ”Doc2Graph” for multi-lingual purposes. The new dataset ”XFUND” [16] is an extension of ”FUNSD” with forms coming from seven different languages, and it would be interesting try self-supervision / continual learning techniques exploiting my library to tackle this task.

2.3 Conferences and Summer Schools

From March 2022 I have been collaborating as a visiting researcher with the Computer Vision Center (CVC), Barcelona. In addition, I have attended the following conferences and summer school:

- Ellis Machine Learning Summer School (on-site), Cambridge University, 11-15 Jul 2022.

- ICPR 2022 (on-site), 21-25 Aug 2022.

- S+SSPR 2022 (on-site), 26-27 Aug 2022.

- Annual Catalan Meeting on Computer Vision ACMCV (on-site), 19 Sep 2022.

3 Publications

Here I list the publication made as a result of my work, accepted by conferences, workshops and journals:

- ”Doc2Graph: a Task Agnostic Document Understanding Framework based on Graph Neural Networks”, TiE @ ECCV 2022 [7] - Workshop Poster, Tel Aviv (Israel)
- ”Graph neural networks and representation embedding for table extraction in PDF documents”, ICPR 2022 [6] - Conference Poster, Montréal (Canada)
- ”Data augmentation on graphs for table type classification”, S+SSPR 2022 [2] - Workshop Oral, Montréal (Canada)
- ”Automatic generation of scientific papers for data augmentation in document layout analysis”, ANDARDA Special Issue (Patter Recognition Letter) [12] - Journal (under review)

4 Credits

During my Ph.D. I have attended the following exams and complementary skills:

Exams (14/18 CFU)

- Probabilistic Graphical Models(Manfred Jaeger). Oct-Nov 2020, (3 CFU)
- Sequence Learning (Paolo Frasconi). Apr 2021, (3 CFU)
- Memory Networks (Federico Becattini). Apr-May 2021, (1 CFU, no exam held)
- Explainable AI (Paolo Frasconi). May 2021, (3 CFU)
- 4th IAPR SSDA 2021 (on-site), Luleå University of Technology, 23-27 Aug 2021: recent developments in document analysis. (5 CFU)
- Towards Developmental Learning (Marco Gori). 20-27 Jun 2022 (5 CFU, no exam held)
- Ellis Machine Learning Summer School (on-site), Cambridge University, 11-15 Jul 2022. (5 CFU, not registered yet)

Complementary Skills (5/6 CFU)

- Incontri potenziamento competenze trasversali. Nov-Dec 2020, (1.5 CFU).
- Impresa Campus Univi: Final presentation of works, Dec 2020, (0.5 CFU).
- Writing, Publishing, Presenting and Searching Scientific Literature, including Journalology. 26-29 January, (3 CFU).
- Impresa Campus Unifi, first call 2021. Feb-Jul 2021. (no credits got)

References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. “SciBERT: A Pretrained Language Model for Scientific Text”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui et al. Association for Computational Linguistics, 2019, pp. 3613–3618. DOI: 10.18653/v1/D19-1371. URL: https://doi.org/10.18653/v1/D19-1371.

[2] Davide del Bimbo, Andrea Gemelli, and Simone Marinai. “Data augmentation on graphs for table type classification”. In: *Proceedings of S+SSPR* (2022). URL: https://arxiv.org/abs/2208.11210.

[3] Michael Bronstein. *Deep learning on graphs: successes, challenges, and next steps*. https://tinyurl.com/deep-learning-on-graphs.

[4] Michael M. Bronstein et al. “Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges”. In: *CoRR* abs/2104.13478 (2021). arXiv: 2104.13478. URL: https://arxiv.org/abs/2104.13478.

[5] Explosion. *spaCy: Industrial-strength NLP*. https://spacy.io/. 2016.

[6] Andrea Gemelli, Emanuele Vivoli, and Simone Marinai. “Graph neural networks and representation embedding for table extraction in PDF documents”. In: *Proceedings of International Conference on Pattern Recognition (ICPR)* (2022). URL: https://arxiv.org/abs/2208.11203.

[7] Andrea Gemelli et al. “Doc2Graph: a Task Agnostic Document Understanding Framework based on Graph Neural Networks”. In: *Proceedings of Text in Everything European Conference on Computer Vision (ECCV)* (2022). URL: https://arxiv.org/abs/2208.11168.

[8] Khurram Azeem Hashmi et al. “Current Status and Performance Analysis of Table Recognition in Document Images with Deep Neural Networks”. In: *CoRR* abs/2104.14272 (2021). arXiv: 2104.14272. URL: https://arxiv.org/abs/2104.14272.

[9] Marcin Kardas et al. “AxCell: Automatic Extraction of Results from Machine Learning Papers”. In: *CoRR* abs/2004.14356 (2020). arXiv: 2004.14356. URL: https://arxiv.org/abs/2004.14356.

[10] Minghao Li et al. *DocBank: A Benchmark Dataset for Document Layout Analysis*. 2020. arXiv: 2006.01038 [cs.CL].

[11] Birgit Pfitzmann et al. “DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis”. In: *arXiv preprint arXiv:2206.01062* (2022).

[12] Lorenzo Pisaneschi, Andrea Gemelli, and Simone Marinai. “Automatic generation of scientific papers for data augmentation in document layout analysis”. In: *Pattern Recognition Letter, Special Issue on Document Analysis (under review)* (2022).

[13] Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. “Rethinking Table Parsing using Graph Neural Networks”. In: *CoRR* abs/1905.13391 (2019). arXiv: 1905.13391. URL: http://arxiv.org/abs/1905.13391.

[14] Pau Riba et al. “Table Detection in Invoice Documents by Graph Neural Networks”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019, pp. 122–127. DOI: 10.1109/ICDAR.2019.00028.

[15] Franco Scarselli et al. “The Graph Neural Network Model”. In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605.

[16] Yiheng Xu et al. “XFUND: A Benchmark Dataset for Multilingual Visually Rich Form Understanding”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3224. DOI: 10.18653/v1/2022.findings-acl.253. URL: https://aclanthology.org/2022.findings-acl.253.

[17] Peng Zhang et al. “VSR: A Unified Framework for Document Layout Analysis combining Vision, Semantics and Relations”. In: *CoRR* abs/2105.06220 (2021). arXiv: 2105.06220. URL: https://arxiv.org/abs/2105.06220.

[18] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. “PubLayNet: largest dataset ever for document layout analysis”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. Sept. 2019, pp. 1015–1022. DOI: 10.1109/ICDAR.2019.00166.