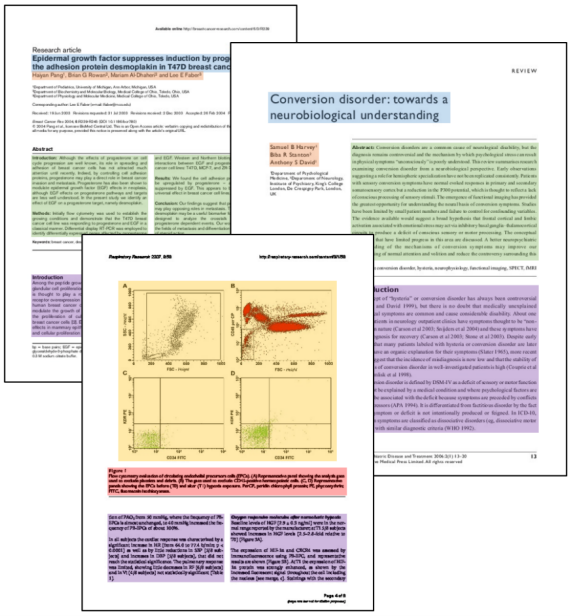
# Multimodal DL methodology for layout identification and classification

Luis Salamanca and Fernando Pérez-Cruz

January 2021




## 1 Project description

Currently, we are carrying out a project on the Swiss parliament archives. This corpus is formed by all the proceedings from 1891 to 1995, from both the National council and the council of states. From these documents we are studying how political interactions, interests, importance of socio-economic factors, etc., have been shaping the taking of political decisions throughout the years. Due to the number of documents and the ample time span, a project of this scale has never been carried out before, which underlines its potential.

Before carrying out any analysis of the text itself, we need to solve a really crucial problem: the extraction of the document layout and the elements within. This problem comprises several other tasks, such as the identification of separate regions of interest, how they are organized in the page, the text contained inside if it exists, the region type, etc. So far, these tasks have been solved on the Swiss parliament data by leveraging on the XML information, that contains both information of the location of different elements in the page and the extracted text. However, this methodology just propagates any errors coming from the original XML, such as wrongly identified bounding boxes, typos in the text, etc.

In the current project, we aim at improving this approach by integrating some recent DL architectures for image and text analysis. In the envisaged methodology, we propose to leverage first on Mask R-CNN [1] to coarsely identify regions of interest, such as paragraphs, lines, tables, figures, etc. To solve this task, we will rely of already existing datasets [2, 4], or generate our own by making use of in-house tools for fast labelling of page regions. Afterward, we propose to tackle the problem of region identification through a hybrid approach that will combine both image and text features, obtained using, for example, R-CNN [5] and LSTM/Transformers [3] architectures respectively. This will not only enable a more precise identification of the region type, but also handling other problems, such as the region ordering, the refinement of the ROIs (e.g. two consecutive regions identified as a table may be a single one), etc. This approach advances the current state-of-the-art by additionally integrating text features, enabling an iterative approach where the masks of the ROIs may be refined according to their content, and possibly, the discovery of regions types from unknown corpora using networks pretrained on different datasets.

## 2 Additional information

- **Difficulty of the project:** From moderate to very challenging.
- **What will you learn?** Natural language processing, recurrent neural networks, convolutional neural networks, language models.
- **Requirements:** Good python level, some knowledge of either Keras or PyTorch, experience with git, machine learning fundamentals, creative thinking.
- **Supervisors:** Luis Salamanca (luis.salamanca@sdsc.ethz.ch) and Fernando Pérez-Cruz (fernando.perezcruz@sdsc.ethz.ch).

## Bibliography

## References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. in proceedings of the ieee international conference on computer vision, 2017. URL: `http://openaccess.thecvf.com/content_ICCV_2017/papers/He_Mask_R-CNN_ICCV_2017_paper.pdf`.

[2] Carlos X Soto and Carlos X Soto. Visual detection with context fo rDocument layout analysis. Technical report, Brookhaven National Lab.(BNL), Upton, NY (United States), 2019. URL: `https://www.aclweb.org/anthology/D19-1348.pdf`.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. URL: `https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

[4] 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)*, pages 1015–1022. IEEE, 2019. URL: https://arxiv.org/pdf/1908.07836.pdf.

[5] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015. URL: https://arxiv.org/abs/1511.08630.