PhD Subject: use of graph attention networks for named entity extraction in scanned documents

Contact: Abdel Belaïd, LORIA lab, READ group, , Nancy, France, abdel.belaid@loria.fr

Starting: January 2, 2019

Motivations

Since the entry into force of the RGPD (General Regulation on Data Protection) on May 25, 2018, European citizens have the legal basis necessary to control and reclaim their personal data. It is a question of providing the technical means that are simple and accessible to as many people as possible to administer their data, to share it, while allowing the organizations that use it to comply with the legislation in force. For this, we want to use massively artificial intelligence to provide a disruptive approach to the management of user data and especially their documents. It is a question of proposing a totally automatic management of the documents of the users, as well at the level of the classification as of the exploitation of the information which they contain. This information (metadata) will then be integrated into an ontology of personal data in order to interrogate these data in a structured way but above all to propose automated reasoning algorithms.

Subject

A first solution has been proposed for the extraction of named entities in personal documents of the "Invoice" type. Documents are scanned, and named entities are extracted from images after Optical Character Recognition (OCR) reading. These entities correspond to the information relating to the issuing company, the delivery and billing addresses, the products ordered as well as the different prices and totals. A deep learning technique based on convolutional network graphs (GCN) has been implemented for the extraction of these entities [4]. It consists in taking advantage of the contextual links in the

neighborhoods of the words of the document. Each node of the graph represents a word of the document and its relations connect it to its four closest neighbors in the four cardinal directions. For the purposes of convolution, the graph in Euclidean space is replaced by the adjacency matrix describing the words in the spectral domain. Thus, the model learns to individually label the words of the document based on their neighborhood links.

In this project, the READ team seeks to extend the extraction method to other types of documents in order to test its ability to extract named entities of different types and in different types of media. We want to test and compare graph attention networks (GAN) of [4, 5, 6, 7, 8] and see which is best suited to the problem. The aim is to better specialize neighborhood graphs to better represent semi-structured information, by retaining the nodes around the headers or information indicators, because it will be tedious to use all the words in the document. This was the case for invoices that are very structured, which is not the case for all documents.

Framework

The implementation will be done in Python using the Keras API based on the Tensorflow library.

References

[1] Devashish Lohani, Abdel Belaid, Yolande Belaid. An Invoice Reading System Using a Graph Convolutional Network International Workshop on Robust Reading, Dec 2018, PERTH, Australia. 2018.

[2] Schlichtkrull, M., T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling 2018. Modeling Relational Data with Graph Convolutional Networks. In Lecture Notes in Computer Science (including subseries Reading, Notes in Artificial Intelligence and Reading Notes in Bioinformatics), volume 10843 LNCS, pp. 593-607.

[3] Jianan Li, Jimei Yang, Aaron Hertzmann, Zhang Jianming, Tingfa Xu, Layoutgan: Generating Graphic Layouts with Wireframe Discriminators, ICLR 2019, pp. 1-16.

[4] Velickovic, P., G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio 2017. Graph Attention Networks.

[5] Gong, L., and Q. Cheng, 2018. Adaptive Edge Features Guided Graph Attention Networks.

[6] Zhang, J., X. Shi, J. Xie, H. Ma, I. King, and D.-Y. Yeung 2018. GaAn: Gated Attention Networks for Learning on Large and Spatiotemporal Graphs.

[7] Monti, F., O. Shchur, A. Bojchevski, O. Litany, S. Günnemann, and M. M. Bronstein 2018. Dual-Primal Graph Convolutional Networks. Pp. 1 to 11.

[8] John Boaz Lee, Ryan A. Rossi, Kim Sungchul, Nesreen Kamel Ahmed, Eunyee Koh, 2018. Attention Models in Graphs: A Survey. 0 (1).