How Do Large Language Models (LLMs) Impact Document Image Analysis? An Open Discussion.

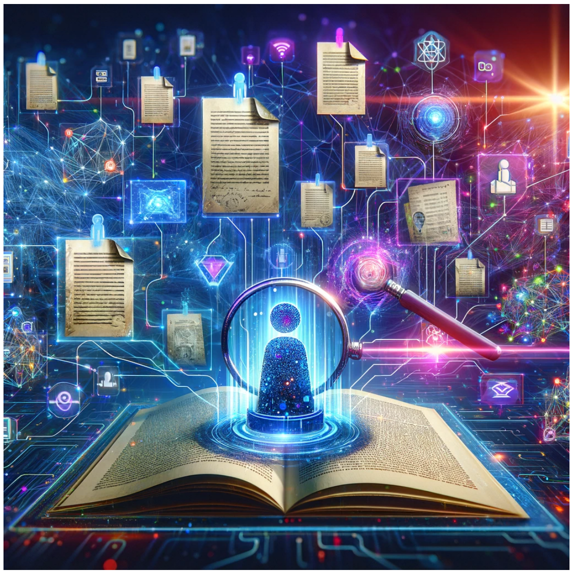

Image created by DALL·E with the prompt: "Hello, can you create an image that shows the connection of large language models with document image analysis?"

In just a few years, Large Language Models (LLMs) have made impressive progress in the areas of artificial intelligence and pattern recognition. Many research fields have been impacted, including finance for price prediction, medicine for molecule creation, and education for personalized learning, to name just a few [10].

LLMs are mostly transformer-based machine learning systems built to generate, understand, and interpret natural language [4]. They are characterized by their large amount of parameters and depend on extensive training data. They achieve state-of-the-art performance in various natural language processing tasks, including translation, summarization, question-answering, and content generation. By analyzing the context and nuances of language, LLMs can produce human-like text, making them valuable for applications that require language understanding and creativity, such as chatbots, content creation, and information retrieval.

Document image analysis is a classical branch of pattern recognition, which aims at recognizing texts and graphics in order to understand written human communication. LLMs already had a significant impact in this field as well in recent years, notably for Optical Character Recognition (OCR) [3], information retrieval [11], and visual question answering [9].

Going beyond purely textual input, current research aims to integrate image and layout aspects as well in the reasoning of LLMs [6], for example when interpreting tables [5]. OCR-free approaches that use LLMs have been developed to perform text reading tasks [7, 8], yet one important drawback of those approaches is their requirement for high-resolution images, which bears an important computational cost [2].

LLMs seems to be able to contribute at all levels of information retrieval. Starting with rewriting the user's query, then searching for the answer, sorting the answers, and finally presenting the results. LLMs can also perform multiple information retrieval tasks simultaneously [11]. In particular, these advances can contribute to the digital humanities. One example is the use of LLMs to support historical research thanks to a conversational interaction with a corpus of document [1].

Despite great performances, conducting scientific experiments with such technology is challenging. First, it involves networks with millions or even billions of parameters. Fine-tuning them is not always possible. Then, when answers are obtained, they are not always in a valid and usable format. This raises the issue of prompt engineering, which, depending on the task, cannot be neglected. Another question is: How can we work with the state of the art when it changes every month if not every week?

Through this presentation, we would like to open a discussion on the impact of LLMs for document image analysis. We will start by describing the technology behind LLMs and present some examples of their use. Then we will outline some limitations of these networks.

References

[1] Giselle Gonzalez Garcia and Christian Weilbach. If the sources could talk: Evaluating large language models for research assistance in history. *arXiv preprint arXiv:2310.10808*, 2023.

[2] Nidhi Hegde, Sujoy Paul, Gagan Madan, and Gaurav Aggarwal. Analyzing the efficacy of an llm-only approach for image-based document question answering. *arXiv preprint arXiv:2309.14389*, 2023.

[3] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

[4] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[5] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Gpt4table: Can large language models understand structured table data? a benchmark and empirical study. https://browse.arxiv.org/pdf/2305.13062.pdf, 2023.

[6] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. https://arxiv.org/pdf/2401.00908.pdf, 2023.

[7] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding. https://arxiv.org/pdf/2307.02499.pdf, 2023.

[8] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.

[9] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. https://arxiv.org/pdf/2306.17107.pdf, 2023.

[10] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[11] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey, 2024.

Author's CV

Anna Scius-Bertrand defended her Ph.D. in Computer Science in 2022 at Ecole des Hautes Etudes in Paris, France. She is currently a PostDoc at the University of Fribourg and the University of Applied Sciences and Arts Western Switzerland (HES-SO, HEIA Fribourg). Her research is focused on handwriting recognition, keyword spotting, and deep learning. She is particularly interested in annotation-free challenges, exploring hybrid machine learning, self-learning, and transfer learning.

Lars Voegtlin started his Ph.D. at the University of Fribourg in the fall of 2018 after finishing in the same year his Master in Computer Science. In his thesis, he presents a deep-learning framework that was created to conduct historical document image analysis experiments. His research interests are self-supervised learning, network initialization, and deep-learning frameworks for historical document image analysis.

Nathan Wegmann is an undergraduate student from the University of Fribourg. He is finishing his BA in Computer Science and Philosophy. He is a Student Assistant in the Department of Informatics. His research interests focuses on the intersection between Philosophy and Computer Science.

Atefeh Fakhari finished her Master's in Computer Science at the University of Fribourg in 2023, where she focused on using graph neural networks to analyze colorectal cancer images for her thesis. Currently, she works as a scientific collaborator at the University of Applied Sciences and Arts Western Switzerland (HES-SO, HEIA Fribourg), continuing her research interests in deep learning frameworks.

Andreas Fischer received the M.S. and Ph.D. degrees in Computer Science from the University of Bern, Switzerland, in 2008 and 2012, respectively. Afterward, he conducted postdoctoral research in Montreal, Canada, at the Centre for Pattern Recognition and Machine Intelligence (CENPARMI) and at Polytechnique Montreal. Currently, he is a Full Professor at the University of Applied Sciences and Arts Western Switzerland (HES-SO, HEIA Fribourg) and a Lecturer at the University of Fribourg, Switzerland. Andreas Fischer's research interests include pattern recognition, deep learning, graph-based methods, document analysis, and handwriting recognition. He has published over 100 peer-reviewed articles in international journals and conference proceedings on these topics. Andreas Fischer is a member of the governing board of the International Association of Pattern Recognition (IAPR), where he represents Switzerland, and Chair of the IAPR technical committee on reading systems (TC11).