Learning tensor-based representations of HTML documents

Context

Information Retrieval (IR) models aim at predicting which documents within a potentially huge collection are relevant to a given user information need (usually a query). Current models of Information Retrieval, like in many other fields, are nowadays based on transformer architectures.
Current research focuses on how to (pre)train the models and the problem of modeling the task better, i.e., how to compute the representation of the document and/or the query, or of both the query and document. Improving the quality of the representation is key to building successful (transformer) models for IR, as shown in the best-performing models to date [1].

Objectives

The internship will explore new ways to compute the representation of (Web) documents. In particular, we aim at exploring more structured representations of texts using tensors. Tensor-based representation is interesting because it allows encoding the information in an efficient manner [2]. This can be interesting to allow retrieving a document among millions efficiently. Tensor-based representations are also more expressive as shown in recent models for math problem solving [3], or abstractive summarization [4].
In the context of Web search, when dealing with web pages, the Document Object Model (DOM) tree represents the document's structure [5]. Recent work on transformer-based models shows that this structure can be encoded explicitly [6] or implicitly [7] in the model. One of the goals of this internship will be to study how tensor-based representation can encode such structures, and assess their effectiveness for large-scale document retrieval.

Organization

The internship will occur at the Qwant offices with visits to the LIP6 (remote work is also possible). It will be supervised by Benjamin Piwowarski from the LIP6 and Lara Perinetti, Romain Deveaud, and Hicham Randrianarivo from Qwant.

The intern will potentially work with the following tools/technologies:

- Deep Learning libraries (PyTorch, TensorFlow, Jax/Flax, Transformers, etc.)
- Python
- Search engine tools (https://github.com/vespa-engine/pyvespa)
- Git repository (Github)
- Jupyter Environment

Qwant will provide the intern a laptop and access to a remote compute server with GPU capabilities.

Candidates can send their resumes to h.randrianarivo@qwant.com, l.perinetti@qwant.com, and bj@piwowarski.fr

Bibliographie

[1]L. Gao and J. Callan, “Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval,” arXiv:2108.05540 [cs], Aug. 2021 [Online]. Available: http://arxiv.org/abs/2108.05540 .

[2]A. Panahi, S. Saeedi, and T. Arodz, “word2ket: Space-efficient Word Embeddings inspired by Quantum Entanglement,” arXiv:1911.04975 [cs, stat], Mar. 2020 [Online]. Available: http://arxiv.org/abs/1911.04975 .

[3]I. Schlag, P. Smolensky, R. Fernandez, N. Jojic, J. Schmidhuber, and J. Gao, “Enhancing the Transformer with Explicit Relational Encoding for Math Problem Solving,” arXiv:1910.06611 [cs, stat], Nov. 2020 [Online]. Available: http://arxiv.org/abs/1910.06611 .

[4]Y. Jiang et al., “Enriching Transformers with Structured Tensor-Product Representations for Abstractive Summarization,” arXiv:2106.01317 [cs], Jun. 2021 [Online]. Available: http://arxiv.org/abs/2106.01317 .

[5]S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, “DOM-based content extraction of HTML documents,” in Proceedings of the twelfth international conference on World Wide Web - WWW '03, Budapest, Hungary, 2003, p. 207, doi: 10.1145/775152.775182 [Online]. Available: http://portal.acm.org/citation.cfm?doid=775152.775182 .

[6]J. Ainslie et al., “ETC: Encoding Long and Structured Inputs in Transformers,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020, pp. 268–284, doi: 10.18653/v1/2020.emnlp-main.19 [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.19 .

[7] Aghajanyan, Armen, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. “HTLM: Hyper-Text Pre-Training and Prompting of Language Models.” ArXiv:2107.06955 [Cs], July 14, 2021 [Online]. Available: http://arxiv.org/abs/2107.06955.