DD2424 - Project Proposal

Michel Le Dez, Pierre Falconnier, Karim Chakroun - Group 50

May 3, 2024

We plan to investigate NLP through text generation models:

- Project to be completed : Default 3 + extensions
  - Upgraded input tokens
  - Replace the RNN architecture with a Transformer
- Title: Enhancing Text Generation Models, From Vanilla RNN to Deep LSTM and Transformers
- Brief description : the goal of the project is to construct and train models which are able to predict the next character in a sentence, compare their performance and investigate improvement strategies. We will rely mostly on the references provided in the project description, [5] for nucleus sampling and resources such as HuggingFace blog about Byte-Pair Encoding and the original papers about BERT[4] and GPT-3 [3] for the Transformer extension.
- Data : we will begin with the Shakespeare dataset [1], and might find other datasets for the extensions
- Software : Pytorch
- How much we will implement: as much as possible to improve our skills in programming for NLP
- Initial set of experiments to run and baseline : prediction of the next characters (similar to the given notebook)
- Milestones:
  - E: Default project 3
  - A/B: the first two extensions (upgrading input tokens and replacing RNN architecture with a transformer)
- Skills/knowledge we aim to acquire (same objectives for all members): RNN, LSTM, NLP (Byte-Pair Encoding, word embedding), Transformers, Transfer Learning.
- Grade we are aiming for : A/B

If we succeed in these previous steps, we would like to explore fine tuning an existing open source LLM like Mistral 7 [2] and compare with our previous results. This could be an opportunity to experiment with tranfer learning.

References

[1] shakespeare. URL: https://www.kaggle.com/datasets/mruanova/shakespeare.

[2] Mistral AI. Mistral 7b. Section: news. URL: https://mistral.ai/news/announcing-mistral-7b/.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. arXiv:2005.14165.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[5] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. arXiv:1904.09751.