Laboratory 2 - Large Language Model

Dr. Laura E. Raileanu, Cédric Campos Carvalho, Elena Najdenovska

15th March, 2024

1 Introduction

1.1 Large Language Model

We have all heard of ChatGPT, but what is it actually? ChatGPT is a chatbot developed by OpenAI and is based on a Large Language Model (LLM). By chatbot, it means a software application (usually a web interface) designed to mimic a human conversation. These models can generate or understand general-purpose language tasks. They are artificial neural networks based on different architectures (*Recurrent Neural Network*, *Transformer*, etc.). Models like these can simply be trained to predict the next word (i.e. GPT-2), but as they are trained in such large quantities of data, they reveal to be useful in many other cases. [1]

The expensive training costs of LLMs, driven by massive data, extreme computational demands, and billions of parameters, pose a significant barrier to wider adoption. Due to these reasons, LLMs are often proprietary and not available under open-source licenses. A noteworthy exception to this trend is Large Language Model Meta AI (LLaMA), an LLM developed by Meta that has been made publicly accessible in four sizes trained with 7, 13, 33, and 65 billion parameters. This open-source release represents a significant contribution to the research community and enables a broader exploration of LLM capabilities and limitations. In their paper, they compare the results of the 13B LLaMA model to the GPT-3 model (with 175B parameters) and tend to obtain better performances on most NLP benchmark tests [2].

A more recent model has been made accessible in Open Source is *Gemma*[1], a family of lightweight, state-of-the-art open models built with the same research and technology than Gemini models developped by Google.

For this Laboratory, we're going to use LLaMA, as it has already multiple libraries that will help to facilitate the use of the model. For example, there's Ollama, an application that will help us to run models locally.

1.2 Retrieval-Augmented Generation

Retrieval Augmented Generation (RAG) is a concept that solves the problem of not having a tuned Large Language Model for your own data. "Real" fine-tuning these models would take lots of computing power and it would be difficult to evaluate it to know if the data is really taken in account. This is why RAG exists by giving access this new data to the LLMs without "refitting" all the parameters.

[1] Visit Google blog post for more information.

2 Setup

To complete the required tasks within this laboratory, it is preferably to use **Python version 3.11**. First of all, create a new environment and install all modules from the `requirements.txt` file.

The simpliest way to install Ollama is with a docker image by executing the following command :

```
docker run -d -v ollama:/root/.ollama -p 11434:11434 \
    --name ollama ollama/ollama
```

If you want to run on GPU to obtain better performances, please follow the instructions in the Github repository.

3 Objective

In this laboratory, you receive two files that you need to complete. The main file, named `notebook.ipynb`, is a Jupyter Notebook that will guide you through all the exercises that you need to complete. Please read **all** the cells as they are giving you instructions that should be followed at each step. The second file named `chat.py` is used as a template for the last exercise that is also described in the main notebook file.

There's a total of three different exercises :

- "*Meet Ollama !*", an introduction to the library *Ollama* and how embeddings work.
- "*Retrieval Agumented Generation (RAG)*", where you explore how to perform RAG on a model with a new source of data.
- "*Create a chat !*", where you use the library `chainlit` to create a chatbot with RAG for an uploaded file.

You need to answer at all questions described in the main file and fill the missing parts of the chat one. Then, upload **both** files in Moodle, the deadline is **27th March, 2024 at 23:59:59**.

No report is needed as all questions can be answered directly in the notebook file.

References

[1] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: 2019. URL: `https://api.semanticscholar.org/CorpusID:160025533`.

[2] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models.* 2023. DOI: `10.48550/ARXIV.2302.13971`. URL: `https://arxiv.org/abs/2302.13971`.