Contribution ID: **26** Type: **Demonstrations & Tutorials**

Leveraging MLflow for Efficient Evaluation and Deployment of Large Language Models

Thursday, 3 October 2024 12:30 (30 minutes)

In recent years, Large Language Models (LLMs) have become powerful tools in the machine learning (ML) field, including features of natural language processing (NLP) and code generation. The employment of these tools often faces complex processes, starting from interacting with a variety of providers to fine-tuning models of a certain degree of appropriateness to meet the project's needs.
This work explores in detail using MLflow [1] in deploying and evaluating two notable LLMs: Mixtral[2] from MistralAI and Databricks Rex (DBRX) [3] from Databricks, both available as open-source models in the HuggingFace portal. The focus lies on enhancing inference efficiency, specifically emphasising the fact that DBRX has better throughput than traditional models of similar scale.
Hence, MLflow offers a unified interface for interacting with various LLM providers through the Deployments Server (previously known as "MLflow AI Gateway") [4], which streamlines the deployment process. Further, with standardised evaluation metrics, we present a comparative analysis between Mixtral and DBRX.
MLflow's LLM Evaluation tools are designed to address the unique challenges of evaluating LLMs. Unlike traditional models, LLMs often lack a single ground truth, making their evaluation more complex.
MLflow allows customers to use a bundle of tools and features that are specifically tailored to deal with difficulties arising from integrating LLMs in a comprehensive manner. The MLflow Deployments Server serves as the central location, eliminating the need to juggle multiple provider APIs and simplifying integration with self-hosted models.
We plan to implement this solution using the MLflow tracking server deployed in the AI4eosc project [5] as a showcase.
In conclusion, this contribution seeks to offer insights into the efficient deployment and evaluation of LLMs using MLflow, with a focus on optimising inference efficiency through a unified user interface. With MLflow capabilities, developers and data scientists can navigate through integrating LLMs into their applications easily and effectively, unlocking their maximum potential for revolutionary AI-driven solutions.

[1] https://mlflow.org
[2] https://huggingface.co/mistralai
[3] https://huggingface.co/databricks
[4] https://mlflow.org/docs/latest/llms/index.html
[5] https://ai4eosc.eu

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: BERBERI, Lisana (KIT-G)

Co-authors: Mr ESTEBAN SANCHIS, Borja (Scientific Computing Centre, Karslruhe Institute of Technology); Dr ALIBABAEI, Khadijeh (Scientific Computing Centre, Karslruhe Institute of Technology); Dr KOZLOV, Valentin (Scientific Computing Centre, Karslruhe Institute of Technology)

Presenter: BERBERI, Lisana (KIT-G)

Session Classification: Demonstrations & Posters