LLM Project to Build and Fine Tune a Large Language Model Project Overview

Overview:

This project builds upon the foundation of LLMs by going through the details of the inner workings. Moreover, it shows how to optimize its usage through prompt engineering and fine-tuning techniques such as LoRA.

To give insight, prompt engineering techniques involve crafting specific instructions or queries given to the language model to influence its output. We will try to generate desired outputs(responses) through zero-shot, one-shot, and few-shot inferences.

Fine-tuning a model means to train a pre-trained model on a task to make it adaptable to other applications. It goes into finetuning and something called Parameter Efficient Fine Tuning (PEFT), which is a technique that optimizes the fine-tuning process by focusing on a model's parameters, making it more resource-friendly.

Project also uses the application of Retrieval Augmented Generation (RAG) using OpenAI's GPT-3.5 Turbo, which results in the development of a chatbot for online shopping for knowledge grounding. Using knowledge grounding and RAG together removes responses that have hallucinations and provides responses that are trustworthy. We are able to mitigate these responses by incorporating information from outside sources to validate and support the generated text.

For instance, using the chatbot in the context of an e-commerce environment, knowledge grounding makes sure that the product information, availability, etc, are sourced from a trusted database or platform. This prevents the AI from making up inaccurate responses and instead gives us information from real-world data.

SPECIAL NOTE:

In order for you to run this project you will need to get access to an OpenAI's API key which can be obtained through OpenAi's website. You will need to either get it free or pay some money to use it. Pricing depends on how much the Ai is being used.

What is the purpose of the project?

The main purpose of the project is to apply critical and practical skills in working with LLMs. It covers the fundamental concepts, advanced techniques, and applications of LLMs. Using these skills we can leverage them for tasks like text generations, fine-tuning, and making chatbots that are built from knowledge-grounded applications.

Data Description:

- The data being used in this project is from the knkarthick/dialogsum dataset from the Hugging Face library.
- Dataset includes textual information about various apparel items and products.

Tech Information:

- Language: Python 3.8
- Libraries: Transformers, datasets, torchdata, torch, streamlit, openai, langchain, unstructured, sentence-transformers, chromadb, evaluate, rouge_score, loralib, and peft.

Code Information:

- Full
- Kb
- Llm_app.py
- Peft
- readme.md
- Requirements.txt

1. The full folder has files that are related to full fine tuning for the model.
2. The kb (knowledge base) directory has two text files (apparel_products.txt and paper_products.txt) that serve as the base for the chatbot, that gives us information about apparel and products.
3. Llm_app.py is the streamlit application, which you can use to interact with the chat bot. It is essential to enter your API key in this application.
4. The peft folder has files that contains info to fine tuning the model.
5. readme.md has information to run the project and has information about the version of python, etc.
6. The requirements text file has all lists of the libraries and the versions to run the project.

Written by: Nikhil Reddy Nalabolu
Email: nikhil.nalabolu@yahoo.com