EECS E6893 Big Data Analytics - Fall 2024

Homework Assignment 4: Generative AI

Due Friday, Nov. 22nd, 2024, by 7:00pm

Assignment 4 guidelines:

Only submit one PDF file with screenshots of your code, brief explanation of the code and the results.

Some helpful links for you:

Transformers library: https://pypi.org/project/transformers/

Peft library: https://pypi.org/project/peft/

Trl library: https://pypi.org/project/trl/

LangChain: https://www.langchain.com/

Part 1 Large Language Models

Please refer to this tutorial and finish task 1-2. https://www.labellerr.com/blog/hands-on-with-fine-tuning-llm/

Task 1 (15 pts)

In this part you are going to answer questions based on some basic definitions of Large Language Models (LLMs) first. Then you should try to deploy Falcon-7b on Google Colab (GPU runtime) and get inference from the model.

Q1.1 Respond with insights grounded in the terminology of Large Language Models. (10 pts)

(1). Can you provide a high-level overview of Transformers' architecture?

(2). What are the two approaches for evaluating language models in NLP, providing brief descriptions of each method along with highlighting their key distinctions?

(3). What is a token in the Large Language Models?

Q1.2 Read through the tutorial slides and deploy Llama 2 on Google Colab and get inference from the model. (5 pts)

(1). Provide screenshots of the results after you successfully download the model and see the text generated.

(2). Change the “max_length” variable in pipline and observe the difference.

Task 2 (40 pts)

In this part you are going to fine tune Llama 2 models based on OpenAssistant dataset.

Q2.1 Write comments for each line of code and succinctly explain what it is doing.

Q2.2 Make a training loss plot.

Q2.3 Use the text generation pipeline to ask questions like “What is a large language model?”

Q2.4 Store fine-tuning Llama2 Model and push Model to your Hugging Face Hub. Provide screenshot of your Hugging Face model page.

Task 3 (25 pts)

In this section, your objective is to leverage the fine-tuning model from Task 2 to construct a versatile chatbot utilizing LangChain. Read through this tutorial: https://python.langchain.com/docs/tutorials/chatbot/

Q3.1, Provide screenshots of the prompt template you have devised.

Q3.2. Provide the text generation outcomes achieved through your chatbot.

Part 2 GPU Coding Experience

Task 4 (20 pts)

Q4.1. Go through this tutorial: https://colab.research.google.com/github/NVIDIA/accelerated-computing-hub/blob/main/gpu-python-tutorial/2.0_Numba.ipynb

Then finish Numba Lab - 1: https://colab.research.google.com/github/NVIDIA/accelerated-computing-hub/blob/main/gpu-python-tutorial/2.1_Numba_lab.ipynb (10 pts)

A note on the Fizz Buzz Problem:

- For each number in the range, if it’s divisible by 3, it’s labeled as “Fizz.”
- If it’s divisible by 5, it’s labeled as “Buzz.”
- If it’s divisible by both 3 and 5, it’s labeled as “Fizz Buzz.”

Q4.2. Go through this tutorial: https://colab.research.google.com/github/NVIDIA/accelerated-computing-hub/blob/main/gpu-python-tutorial/3.0_Numba_gauss.ipynb

Then finish Numba Lab - 2: https://colab.research.google.com/github/NVIDIA/accelerated-computing-hub/blob/main/gpu-python-tutorial/3.1_Numba_lab_2.ipynb (10 pts)

Take screenshots of your code and results.