Progress Project Report

Members
Member 1: Tran Khanh Bang (V202100411)
Member 2: Truong Gia Bao (V202100564)

Date
August 21, 2024
Advisor: Prof. Wray Buntine

1 Introduction

This report serves to document the ongoing progress and developments of the Chatbot project designed specifically for the Computer Science (CS) students at VinUniversity. The primary aim of this report is to provide a comprehensive update to stakeholders on the current status of the project, aligning with the final goals and demonstrating adherence to the outlined academic and practical objectives.

2 Challenges

Our project faces these major challenges:

- Leveraging OpenAI's API or undertaking fine-tuning on our infrastructure necessitates significant computational resources, which are currently beyond our project's budget. Additionally, the time required for fine-tuning these models is considerable, potentially impacting project timelines.
- Limited server resources, including insufficient disk space, hindered the installation of necessary packages and the deployment of the final model.
- While the CECS server provided GPU acceleration for model fine-tuning, limitations in deploying the large model to GitHub due to size constraints and LFS package issues posed additional obstacles (only the paid version can use more than 1GB)

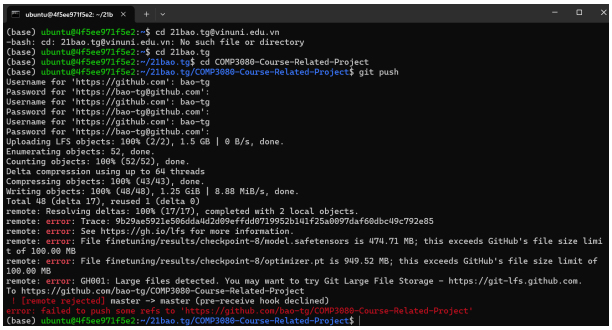

Figure 1: LFS package issues of GitHub

- The inability to run the user interface locally on the CECS server using Streamlit impacted the development and testing process.

Figure 2: Streamlit run locally

3 Current implementation

Our implementation strategy prioritizes cost-effectiveness and accessibility. To achieve this, we utilize a combination of open-source resources, the implementation in details can be found in the footnote[1]:

- **Fine-Tuning Dataset:** We leverage the "ArtifactAI/arxiv-beir-cs-ml-generated-queries" dataset available on Hugging Face (huggingface.co). This pre-existing dataset, specifically focused on computer science and machine learning, provides a strong foundation for fine-tuning the chosen language model.

Figure 3: Finetuning 1000 data, 8 batch, 1 epoch(toy example)

- **Language Model and Tokenizer:** We employ the widely used GPT-2 model ("model = GPT2LMHeadModel.from_pretrained('gpt2'), tokenizer = GPT2Tokenizer.from_pretrained('gpt2')") from Hugging Face. This combination offers a good balance between performance and resource efficiency, making it suitable for deployment with Streamlit.
- **High-Performance Fine-Tuning:** To expedite the fine-tuning process, we leverage the CECS server's robust infrastructure, including 6 NVIDIA A5000 GPUs, 64 CPUs, and 230GB RAM. This computational power enables efficient training of the language model on our dataset, accelerating development timelines.
- **User Interface Development:** The user interface of the chatbot is developed using `Streamlit`, we also found the way to deploy the model as an application in the Streamlit repository. This choice was made to ensure that the interface is not only intuitive and user-friendly but also capable of supporting the dynamic requirements of interactive chatbot functionalities.

4 Teamwork

- Truong Gia Bao (Project Manager - Software Engineer):
  - Led the project planning and coordination efforts.
  - Oversaw the overall project timeline and ensured timely delivery.
  - Developed and maintained core system functionalities.
- Tran Khanh Bang (Software Engineer):
  - Contributed to project development by implementing features and fixing bugs.
  - Collaborated with the team to ensure successful app delivery.
  - Demonstrated technical expertise in connecting the UI components.

5 Goals in the final phase

To further optimize the chatbot's capabilities and user experience, we plan to implement several enhancements:

- **Dataset Expansion:** We will significantly expand the training dataset by incorporating a wider range of CS-related materials, including textbooks, research papers, and code repositories. This will enrich the chatbot's knowledge base and improve its ability to provide comprehensive and informative responses.

[1] See here.

- **Model Upgrading:** The current model size may limit the chatbot's capacity for complex reasoning and context understanding. We aim to explore larger language models to enhance these capabilities, potentially leading to more sophisticated and nuanced interactions.
- **Resolution of the LFS Package Issues on GitHub:** In the final phase, the objective is to address and resolve the size constraint limitations of the LFS package.

6 Conclusion

This report outlines the development of a CS student chatbot, addressing challenges related to resource constraints and model performance. By leveraging open-source tools, high-performance computing, and a user-centric design, we have created a foundation for a valuable educational resource. Future enhancements will focus on expanding the dataset, upgrading the language model, and implementing user data management to further improve the chatbot's capabilities.

7 Appendices

1. Hugging Face. (n.d.). ArtifactAI/arxiv-beir-cs-ml-generated-queries [Dataset]. Retrieved from https://huggingface.co/datasets/ArtifactAI/arxiv-beir-cs-ml-generated-queries/tree/main
2. Hugging Face. (n.d.). transformers: Pre-trained Deep Learning Models for NLP [Software]. Retrieved from https://huggingface.co/transformers
3. Streamlit. (n.d.). *Streamlit Official Website*. Retrieved from https://streamlit.io/