Final Project

Local LLM for Mac Silicon Devices

CSCI S-104 Advanced Deep Learning

Professor Zoran B. Djordjević
Name: Gassan Yacteen
Date: May 08, 2024

Abstract

In this report, I present the outcomes of my final project for the Advanced Deep Learning course, focusing on the development and optimization of a local machine learning framework designed to run large language models (LLMs) on Apple Silicon devices. This initiative was primarily driven by the need for enhanced runtime efficiency and the creation of digital tools that prioritize user privacy, a pressing concern for macOS users without access to traditional GPU acceleration technologies like CUDA.

The project began by employing LocalGPT and SiLLM, tools aimed at operating efficiently on Apple hardware while securing data processing locally to bolster privacy. However, these tools faced substantial practical challenges due to their inefficiencies and performance limitations on Apple Silicon. This challenge necessitated a strategic shift to using the mlx library, which aligned better with Apple's Metal API and effectively harnessed the native hardware acceleration of an Apple Mac M2 Pro chip equipped with 16 GPUs, 16 CPUs, and 16GB of RAM.

This hardware setup enabled the successful local training of the Llama3 Instruct model, subsequently fine-tuned on a mathematics dataset. The project not only showcased the feasibility of using locally executed machine learning frameworks on non-traditional hardware but also led to the implementation of a private LLM chatbot. This chatbot, designed to operate entirely within the local environment, exemplifies the project's aim to develop privacy-centric AI applications on Apple Silicon.

Throughout this endeavor, challenges such as hardware limitations were prominent, limiting the extent of model training possible without further system upgrades. However, this project was a significant learning opportunity, providing deep insights into the complexities of local LLM deployment and machine learning workflow optimization on specialized hardware platforms.

- YouTube Brief Overview: https://youtu.be/JIO05ND9Mpk

- YouTube Detailed Walkthrough: https://youtu.be/ujRlI79gBS8

This summary aims to encapsulate the project's scope, including the problems addressed, technologies utilized, practical applications, benefits, drawbacks, and the educational experiences derived from implementing a secure and private LLM chatbot. It serves as a basic overview for colleagues to assess the relevance and potential of the documentation and code available in the full report.