Imperial College London

MEng Individual Project

Department of Computing

Imperial College of Science, Technology and Medicine

A New Scalable Runtime for LLM Inference

Author:
Hamish McCreanor

Supervisor:
Peter Pietzuch

January 14, 2025

Submitted in partial fulfillment of the requirements for the MEng Computing of Imperial College London

Contents

Chapter 1

Introduction

As large language models (LLMs) are found useful for ever-wider classes of applications, a trend has arisen focusing on the low-cost, local deployment of these systems. While the training of LLMs like LLaMA, BERT and OpenAI's GPTs typically requires months of training and is prohibitive for all but the most well-funded of organisations, performing inference on these models locally is comparatively more feasible. This enables developers to create services with tighter LLM integrations - instead of calling a black-box API provided by an LLM provider, they can instead run a local version of the LLM, tuning the inference runtime to more appropriately match the context in which it is called.

As a result, there is currently a vast body of research aiming to improve existing inference systems. The aim of this is to improve LLM inference performance along various axes. These include running on lower-powered hardware; running with improved throughput and running at greater energy efficiencies. These optimisations focus on specific elements of the inference pipeline, particularly improving KV cache usage and kernel fusion. To build systems containing these optimisations, developers frequently turn to high level languages like Python in order to quickly develop the infrastructure surrounding their new technique. Developing this way limits the ability of the system to exploit memory-access patterns and application parallelism (especially in a language like Python, with its global interpreter lock) and incurs unnecessary overhead.

This project aims to build on the existing llama.cpp inference server (see 2.2.5) to deliver a system that improves the dispatch of compute kernels by better parallelising the inference pipeline. Hello [1]

Chapter 2

Background

2.1 Preliminaries

2.1.1 LLM Architecture

Transformer Architecture

2.1.2 LLM Inference

KV Cache

Parallelism

Request Batching

2.2 Related Work

2.2.1 vLLM

PagedAttention

2.2.2 Triton

2.2.3 SGLang

2.2.4 Triton

2.2.5 llama.cpp

Chapter 3

Project Plan

Chapter 4

Evaluation Plan

4.1 Functional Requirements

4.2 Performance Metrics

Chapter 5

Ethical Issues

The principal two ethical concerns of a project in this field relate to the potential for misuse as well as provenance issues surrounding the dataset on which the model was trained.

The potential for misuse of LLMs is vast, with many instances of LLM abuse already being documented. LLM abuse typically involves the use of the model to produce harmful or misleading content. There already exist instances of LLMs being used to generate phishing messages, with the intent to produce emails that sound more plausible and are more likely to be engaged with by a target. In addition to this, LLMs can be used to produce vast quantities misinformation or biased content that are then published to social media platforms. End users may be unable to distinguish between content created by a genuine user and content generated by an LLM and thus end up misinformed.

The large size of the datasets required to train these models create potential ethical and data protection issues. Concerns exist regarding the ability for generative models to amplify existing biases in their training data, with some of these concerns borne out in cases like Microsoft's Tay chatbot. AI fairness is still an open area of research and it is unlikely that existing LLM models will be completely free of bias at inference time. At the same time, the provenance of this training data is also an important ethical consideration. Private or sensitive data has the potential to be incorporated into training sets and there exist cases where this training data has then been generated verbatim at inference time, exposing this sensitive data to an end user.

If successful, our project broadens access to LLMs by making better use of available hardware to perform inference. This increases the viability of local inference and opens up these models to a greater proportion of hardware configurations and thus a greater number of users. While this represents a boon for the accessibility of this technology, with users no longer limited to a handful of offerings by large companies, it also increases the number of potentially malicious actors who are able to use LLMs. Small and local deployments likely have less of the oversight that large LLM providers experience, and thus are more able to misuse this technology. These two elements must be carefully managed in order to produce a project that adheres to reasonable ethical standards.

As this project represents a proof-of-concept, rather than a full-featured inference

engine, any advances made are unlikely to immediately be adopted and thus any ethical concerns are likely to be uncovered at a pace with which they can be identified early on and mitigated quickly.

Chapter 6

Bibliography

[1] Einstein A. Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]. Annalen der Physik. 1905;322(10):891-921. pages 2