Towards Efficient and Sustainable LLM Deployments

Zhenhua Liu (Applied Mathematics & Statistics), Anshul Gandhi (Computer Science)

Project Overview

Large Language Models (LLMs) and their remarkable applications, such as ChatGPT, have disrupted the AI landscape in the past few months. For example, ChatGPT recorded 100 million monthly active users in January 2023, after merely two months of its release, making it the fastest growing application [1]. Every major software company, including Google, Microsoft, and Meta, has immediately responded by making substantial monetary and workforce investments in developing their own LLMs and associated applications.

But, this explosive growth in LLMs has resulted in a serious, unintended consequence—uninhibited and unregulated rise in carbon emissions. For example, developing the state-of-the-art GPT-3 model [2] produced as much carbon emissions as 123 gasoline-powered passenger vehicles driven for one year [3]. Given the projected exponential growth in LLM deployments [4], there is an immediate need to develop techniques to reduce the carbon footprint of LLMs. Simply put, if nothing is done, the current growth rate of LLMs is unsustainable. Enterprise companies and industries seek to maximize profit, and so do not have the right incentives in place to prioritize the reduction of carbon emissions. As such, academia can play a vital role in addressing the carbon emissions problem.

Technically, what is urgently needed is a framework to reduce the carbon footprint of not just today's LLMs, but also those that are sure to emerge in the next few months and years. The goal of this project is to develop exactly such a framework. This seed grant proposal lays out our plans to produce convincing preliminary results to demonstrate the feasibility of such a framework. Sustainability and LLMs are both timely topics with funding agencies, such as NSF, dedicating new programs specifically for these areas (e.g., NSF's newly released DESC program [5]). Thus, funding from the seed grant now will enable the PIs to develop a strong proposal submission to external agencies by next year.

Intellectual merit and broader impact. This seed project will develop a *novel optimization framework* to choose the right deployment strategy and parameter settings for any given LLM from a few empirical runs. Specifically, the framework will determine how to model- and data-parallelize the LLM execution over available GPUs to minimize carbon costs will meeting throughput targets. The framework will rely on carbon cost and performance *models* that the PIs will develop based on experiments conducted under diverse deployment and parameter settings.

The project will significantly advance the state-of-the-art by replacing existing "rule-of-thumb" static deployment strategies (e.g., model parallelize on all available GPUs) with an optimized-based framework that automatically determines the best strategy (e.g., model parallelize on a quarter of the GPUs and data parallelize over the setup $4\times$) and parameter settings for an arbitrary LLM. The framework will aid our understanding of how these parallelization strategies impact the carbon cost of LLM deployments. To the best of our knowledge, such a solution framework does not yet exist for LLMs.

Team qualifications. The research team at SBU is uniquely qualified to lead this project. PI Liu (from AMS, affiliate appointment in CS) has significant expertise in the emerging topic of LLMs and AI/ML systems in general [6–19] , as well as sustainable data centers [20–28]. Co-PI Gandhi (from CS, affiliate appointment in AMS) is an expert in sustainable computing [29–45]. Both PIs have recently collaborated successfully on NSF and DOE grants, including their recent $1.5 Million NSF Large grant on sustainability. Further, both PIs have also co-authored publications on efficient deep neural networks and sustainability [35, 46].

Alignment with seed grant objectives. The proposed project on sustainable LLM deployments is a perfect fit for the team's expertise and presents a timely opportunity for the PIs to establish a foothold in this emerging and critical area. What the team is missing is the necessary early funding to support students for obtaining preliminary but convincing results on sustainable LLM deployments. Funding from the seed grant will allow the team to realize these objectives, providing the PIs with a much needed competitive edge for a strong, full proposal on the timely topic of sustainable LLMs.