Towards Greener AI: Predictive Power Modeling and Optimization for Energy-Efficient GPU Computing

Saurabhsingh Rajput (saurabh@dal.ca; +1-782-882-3468),
Faculty of Computer Science, Dalhousie University.

Advisor: Dr. Tushar Sharma (tushar@dal.ca),
Faculty of Computer Science, Dalhousie University.
Category: Clean Tech Innovation - AI, Energy efficiency, Sustainability.

Research overview

Artificial Intelligence (AI)-based applications have become ubiquitous, touching almost all aspects of modern human life. However, the energy-intensive nature of AI models is at odds with the need for low-power, high-performance computing. Large AI models consume substantial energy during training and inference, leading to significant carbon emissions. Green AI practices focus on energy-efficient AI models and technologies to reduce the environmental impact of compute-extensive AI. While there is growing awareness of energy-efficient AI, targeted profiling and benchmarking of AI workloads is the need of the hour. Comprehensive profiling considering factors from hardware devices to system software and software applications is vital to provide data and insights to facilitate impactful Green AI research and real-world optimizations.

Our research develops a predictive power model and optimizer to improve the energy efficiency of AI models on Graphics Processing Units (GPUs), an increasingly essential technology for parallel AI workloads. The methodology involves hands-on profiling of representative GPU workloads, directly measuring the relationships between power consumption, memory usage, and computations. These real-world insights are integrated into a multi-objective power model that minimizes energy usage while meeting performance requirements.

The outcome of this research will be a validated power model and optimization technique to configure GPUs for energy-efficient AI workload execution dynamically. The model profiles real-world GPU workloads to extract fine-grained power-performance insights. These insights enable optimizing power consumption while meeting runtime targets by tuning parameters like GPU frequency based on expected workloads. This study provides a pathway to sustainable AI by enhancing GPU energy proportionality without compromising the performance of the AI models. By open-sourcing this research, we aim to provide the AI community with tools to benchmark and optimize AI's carbon footprint. This research has the potential for real-world impact on reducing emissions for GPU-powered industries globally.