Significantly reducing energy consumption of Deep Neural Networks

PI: Niranjan Balasuramanian, Co-PIs: Aruna Balasubramanian, Anshul Gandhi

Computer Science Department, Stony Brook University (SBU)

Project Abstract

Advances in Deep Neural Networks (DNNs) have led to several wide ranging improvements in AI-based technologies across many domains, resulting in practical applications such as ChatGPT. However, these developments have come at great computational and energy costs. In response, developers and practitioners are starting to explore available model optimizations and hardware choices that will help reduce the energy cost of training and running DNNs. Unfortunately, it is not easy to estimate the energy impact of running a given DNN model under a given optimization across all available hardware. Given the range of choices, it is not feasible to experimentally evaluate the impact of each model, optimization, and hardware choice. Our proposal tackles these problems head-on.

In particular, we propose to design a novel energy prediction algorithm that can accurately predict the energy of running a model (or its component) on any given hardware (even ones not seen during its training). We will also develop advisory tools based on our prediction algorithm to demonstrate the energy savings that can be achieved. Preliminary results from these thrusts will allow the PIs to provide convincing data to support a full proposal to external funding agencies. As part of the full proposal, we will also design an optimization algorithm for a distributed training setup that jointly searches for the best model optimization and model distribution over a given distributed infrastructure.

Intellectual Merit. The proposed framework will fill an important gap in energy-efficient DNN research and application. Successful completion of the proposal will yield: (i) scalable energy prediction of DNN models on *previously unseen hardware* using a novel hardware representation. The representation captures similarities and variations in hardware devices by learning a hardware embedding. (ii) a what-if and bottleneck analysis framework to predict the energy effect of different *DNN optimizations* by designing a tree transformation algorithm that transforms the original model tree to a new model tree (post optimization), and then extrapolates the energy for the new tree, (iii) joint model optimization and distribution techniques for energy efficient distributed training of large DNNs. Taken together, the proposed work will improve understanding of how hardware, optimizations, training strategies, and runtime environments affect DNN energy usage.

Broader Impact. Power consumption is arguably one of the biggest challenges facing the computing industry today. Deep learning applications are driven by increasingly complex algorithms, resulting in significant power consumption [6, 40]. A recent study [40] shows that the estimated CO2 emission to train a Transformer model is more than 5 times the CO2 emission during the lifetime of a car. Our proposal is a step towards designing energy efficient deep learning models, with the potential to make wide-ranging impact including reducing carbon footprint, reducing costs for deploying deep learning applications, and allowing users to run critical applications on their battery-operated devices.

The PIs have successfully transferred research results into practice in their prior work. Gandhi's work on autoscaling [24, 25] was adopted by Facebook on its production clusters [49]. Aruna Balasubramanian works closely with Google and has won a Google Research Award and a Google Chrome Award for her work on improving Web performance. We will leverage these collaborations, and build new ones, to further disseminate our results. Finally, all three PIs have successfully involved high school (HS) students and undergraduates (UG) in their research on NLP, energy consumption, and modeling. This project will allow the PIs to continue this outreach practice.

Fit for Seed Grant. This project will provide the necessary first steps in reigning in the energy use of DNNs—the ability to accurately predict the energy consumption of an arbitrary DNN on an arbitrary hardware. Such predictions can then guide model optimizations and empower DNN developers. Given the project's societal impact and the timely nature of DNN research, obtaining preliminary results via the seed grant will greatly improve the team's chances of putting together a competitive proposal for external funding. Further, the seed grant will enable the PIs to pivot their individual research areas towards the timely topic of energy-efficient DNNs, providing them access to various funding opportunities in the broad area of sustainability [36]. Such funding avenues will aid the career advancement of the PIs.