Performance and Scalability of Distributed Deep Learning

Shirley Moore, University of Texas at El Paso (Principal Investigator)
Deepak Tosh, University of Texas at El Paso (Co-Investigator)
Joshua Suetterlein, Pacific Northwest National Laboratory (Co-Investigator)
Joseph Manzano, Pacific Northwest National Laboratory (Co-Investigator)

The project addresses the objective of intelligent, adaptive resource management and efficient use of heterogeneous computing technologies for parallel and distributed deep learning applications. Deep learning is increasingly being used to augment traditional modeling and simulation in a number of scientific and engineering application areas. While GPUs are currently the main workhorse for parallel deep learning, domain-specific accelerators such as TPUs are claiming some of the market and predicted to grow in importance. Because of increasingly large and complex datasets for which the computational and memory demands exceed the resources of a single CPU or GPU, various forms of parallelism are used to scale up deep learning. While versions of popular deep learning tools implement parallelism and some performance analysis and optimization tools and guidelines exist, parallelization strategies and performance tuning are based largely on empirical and trial-and-error methods rather than model-based methods. Also, there is no standard portable methodology and tool suite to analyze performance of different deep learning frameworks across multiple processor architectures (e.g., different vendor GPUs, TPUs) and communication networks. The goal of this project is to provide the users and developers of deep learning frameworks, as well as system administrators, with the tools needed to analyze, optimize, and scale deep learning models on high performance computing platforms. This project does not address strategies for increasing the accuracy of deep neural network (DNN) models, other than assessing how such strategies affect performance, but rather focuses on how efficiency and scalability of a given approach can be evaluated and optimized. The technical approach described below encompasses performance modeling, communication tracing and analysis, and analysis of coupled machine learning/simulation workflows.

A performance model that illustrates how the DNN is mapped onto the underlying hardware resources can reduce the search space for efficient solutions. The project is developing models for memory consumption, data movement, and communication costs of deep learning applications. Performance models will assist deep learning developers in selecting parallelization strategies and configurations and in determining opportunities for performance optimization.

Developing a unified approach to collecting communication traces will enable a portable cross-platform analysis of communication performance, including communication analysis for coupled simulation and deep learning workflows. The approach is to provide tools for collecting and examining communication traces for different deep learning frameworks on a variety of DNN models. Being able to trace communication for coupled simulation and deep learning workflows with the HPC community accepted Score-P standard will enable integrated performance analysis of these workflows.

Coupling of simulation and deep learning will bring new challenges in resource management, performance optimization, and scalability. A unified instrumentation and analysis framework for the coupled workflows will help address these challenges. The approach is to design a framework for low-overhead performance monitoring with the goal of collecting data about resource utilization, data movement, and excessive wait times so as to be able to identify and fix bottlenecks in a coupled workflow.