Invited Online Talk

Efficient Neural Network Optimization on Supercomputers

Zhao Zhang

Time: Tuesday, 08/01/2023, 10:00 AM — 11:00 AM (CST)

Zoom Link: https://unt.zoom.us/j/85424913940

Abstract: We have seen the fusion of HPC and deep learning (DL) over the past few years: The powerful HPC memory and communication architectures of HPC systems have been shown to support DL applications well; Scientists are exploring and exploiting DL to solve domain research challenges. In this talk, I will share my recent research on scalable second-order optimization with Kronecker-factored Approximate Curvature (K-FAC) for neural networks. I will introduce KAISA, a K-FAC-enabled, Adaptable, Improved, and ScAlable second-order optimizer framework that adapts the memory footprint, communication, and computation given specific models and hardware to improve performance and increase scalability. We quantify the tradeoffs between memory and communication cost and evaluate KAISA on large models, including ResNet-50, Mask R-CNN, U-Net, and BERT, on up to 128 NVIDIA A100 GPUs. Compared to the original optimizers, KAISA converges 18.1-36.3%faster across applications with the same global batch size. Under a fixed memory budget, KAISA converges 32.5% and 41.6% faster in ResNet-50 and BERT-Large, respectively.

Bio: Dr. Zhao Zhang is a computer scientist and leads the scalable computing intelligence group in the Data Intensive Computing Group at Texas Advanced Computing Center. Before joining TACC, he spent two years as a postdoc researcher in AMPLab and BIDS (Berkeley Institute of Data Science) at the University of California, Berkeley, working with Prof. Michael J. Franklin. He received his Ph.D. degree from the Department of Computer Science at the University of Chicago in 2014 under the supervision of Prof. Ian Foster. He will join the Department of Electrical and Computer Engineering at Rutgers University as an assistant professor in Fall 2023. Dr. Zhang has published more than 60 papers, such as SC, HPDC, and TPDS. Also, he has received multiple NSF grants, such as Core and CSSI, and co-leads the cyberinfrastructure research thrust in the NSF ICICLE AI Institute.