UNIVERSITY OF CALIFORNIA MERCED | SCHOOL OF ENGINEERING

Electrical Engineering and Computer Science Technical Seminar Series

Friday, December 13, 2019
12:00 PM in COB 263

Performance Prediction and Optimization of Parallel and Distributed Deep Learning

Dr. Cong Xu

Faculty Host: Prof. Dong Li

Abstract

In this talk, I will first present an predictive performance model for distributed deep learning training. The goal is to answer some key questions including: 1. how fast can a neural network be trained on a specific platform before we build the hardware to run it? 2. what is the optimal system configuration to train a given model? 3. how much speedup can we get if another compute node is added to a cluster or if existing interconnect is upgraded, without running full-scale experiments on target hardware? 4. what are the bottlenecks in distributed training? We used the model to quantize one bottleneck in parallel and distributed deep learning - the high network communication cost for synchronizing. I will talk about our solution named TernGrad to tackle this issue. TernGrad uses ternary gradients, which requires only three numerical levels {-1,0,1}, to aggressively reduce the communication time. I will cover some techniques that we proposed to improve its convergence including layer-wise ternarizing and gradient clipping. Our experiments show that applying TernGrad on variours CNNs does not incur any accuracy loss and can even improve accuracy, while achieving significant speed gains. Finally I will briefly go through some other activities on deep learning in Hewlett Packard Labs.

For additional information contact Prof. Wan Du <wdu3@ucmerced.edu>

Cong Xu
Hewlett Packard Labs

Biography

Cong Xu is currently a research scientist at Hewlett Packard Labs. His research interests include deep learning acceleration, high performance computing, and distributed file system. He has published more than 30 papers on non-volatile memory and deep learning acceleration. He co-led the development of NVSim, which is a widely used simulation framework for non-volatile memories. Cong received his B.S. degree in microelectronics from Peking University, Beijing, in 2009, and his Ph.D. degree in computer science and engineering from Pennsylvania State University in 2015.