Seminar – Tuesday, February 4, in 100A Riggs Hall 11:00 a.m. – 12:00 p.m.

Holcombe Department of Electrical and Computer Engineering Seminar Series

Ultra-High-Performance and Energy-Efficient Deep Learning for Resource-Constrained Devices

Ao Ren

Ph.D. candidate, Dept. of Electrical & Computer Engineering,
Northeastern University

Abstract

Recent successes of deep learning in various domains rely on the excessive provision of weight parameters and the resulting large number of computations, which consume tremendous memory storage and energy. As a result, deep learning models can hardly be deployed onto resource-constrained devices, such as mobile devices and IoTs. In this talk, I will discuss our representative work in addressing this problem. I will discuss our ADMM-NN compression framework, which has achieved the highest compression ratios on diverse representative DNN models. For instance, the parameter reduction on ResNet-50 is 25×, and that on LeNet-5 is 2,000×. Then, I will describe our DARB algorithm, which aims to overcome the limitations brought by conventional irregular pruning. It has significantly outperformed prior structured pruning methods by 4×~9× on multiple RNN models. Besides DNN model compression, we also study the emerging computing technologies for next-generation AI systems. I will describe our SC-DCNN, the first stochastic computing-based DCNN accelerator. Further, we figure out the suitability between stochastic computing and AQFP superconducting technology. I will describe our efforts in proposing the first stochastic computing and AQFP-based DNN accelerator, which has achieved the highest energy efficiency among all the current DNN accelerators. The energy gain is 69,000× compared with its CMOS counterpart. Finally, I will conclude my talk with my on-going research and plan towards energy-efficient AI.

Biography of Speaker

Ao Ren is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering at Northeastern University. His research centers around high-performance and energy-efficient deep learning systems, which include deep learning compression algorithms, accelerator architecture, chip design, and emerging computing technologies. His work has appeared in some of the top venues in computer architecture and artificial intelligence, such as ASPLOS, ISCA, AAAI, ISSCC, and ICCAD. He has also received the Best Paper Award and Best Student Presentation Award in ICCASP’2017.