USC Viterbi School of Engineering

Seminar

Ming Hsieh Department of Electrical and Computer Engineering

Efficient Deep Learning with Sparsity: Algorithms, Systems, and Applications

Zhijian Liu
PhD Candidate
Massachusetts Institute of Technology

Friday, March 8, 2024 | 10:30am PST | EEB 132
Zoom Meeting ID: 967 9033 7008 | Passcode: 406845

Abstract: Machine learning is widely used across a broad spectrum of applications. However, behind its remarkable performance lies an increasing gap between the demand for and supply of computation. On the demand side, the computational costs of machine learning models have surged dramatically, driven by ever-larger input and model sizes. On the supply side, as Moore's Law slows down, hardware no longer delivers increasing performance within the same power budget.

In this talk, I will discuss my research efforts to bridge this demand-supply gap through the lens of sparsity. I will begin by discussing my research on input sparsity. First, I will introduce algorithms that systematically eliminate the least important patches/tokens from dense input data, such as images, enabling up to 60% sparsity without any loss in accuracy. Then, I will present the system library that we have developed to effectively translate the theoretical savings from sparsity to practical speedups on hardware. Our system is up to 3 times faster than the leading industry solution from NVIDIA. Following this, I will touch on my research on model sparsity, highlighting a family of automated, hardware-aware model compression frameworks that surpass manual solutions in accuracy and reduce the design process from weeks of human efforts to mere hours of GPU computation. Finally, I will present several examples demonstrating the use of sparsity to accelerate computation-intensive AI applications, such as autonomous driving, language modeling, and high-energy physics. I will conclude this talk with an overview of my ongoing work and my vision towards building more efficient and accessible AI.

Bio: Zhijian Liu is a Ph.D. candidate at MIT, advised by Song Han. His research focuses on efficient machine learning. He has developed efficient ML algorithms and provided them with effective system/algorithm support. He has also contributed to accelerating computation-intensive AI applications in computer vision, natural language processing, and scientific discovery. His work has been featured as oral and spotlight presentations at conferences such as NeurIPS, ICLR, and CVPR. He was selected as the recipient of the Qualcomm Innovation Fellowship and the NVIDIA Graduate Fellowship. He was also recognized as a Rising Star in ML and Systems by MLCommons and a Rising Star in Data Science by UChicago and UCSD. Previously, he was the founding research scientist at OmniML, which was acquired by NVIDIA.

Hosts: Mahdi Soltanolkotabi, soltanol@usc.edu
Peter Beerel, pabeerel@usc.edu