Invited Online Talk

Memory Efficient Machine Learning Training

Dr. Jie Ren

Time: Wednesday, 05/29/2024, 11:00 AM — 12:00 PM (CST)

Zoom Link: https://unt.zoom.us/j/86512602380

Email: jren03@wm.edu

Abstract: Deep neural networks (DNNs) are becoming increasingly deeper and wider due to the growing demands on prediction accuracy and analysis quality. Training wide and deep neural networks can be extremely memory-consuming. However, state-of-the-art accelerators such as GPUs are only equipped with the limited capacity of high-bandwidth memory due to hardware design constraints, which significantly limits the maximum model and batch size that can be trained. As a result, the bottleneck for state-of-the-art model development is now memory rather than data and compute availability, and we expect this trend to worsen in the future.

This talk will discuss using memory-efficient training techniques in runtime systems for large-scale DNN model training. Specifically, I will show how to train different types of multi-billion parameter models with limited GPU recourses by exploring the usage of heterogeneous memory and extra computation resources on GPUs and CPUs.

Bio: Dr. Ren Jie is an Assistant Professor of Computer Science at William & Mary. She received her Ph.D. from the University of California, Merced, and her B.S. degree from the Beijing Institute of Technology. Her research interests span operating systems, computer architecture, and their intersection with machine learning and high-performance computing. Her research aims to improve the performance and resource efficiency of heterogeneous computing systems while making it easier for users to deploy and manage their applications.