PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering

Faculty Candidate Seminar – Software Engineering

*****CONFIDENTIAL – PLEASE DO NOT POST PUBLICLY*****

Minjia Zhang
Researcher, Microsoft

Monday, January 23, 2023
Presentation: 10:30 A.M. – 11:30 A.M.
Reception: 11:30 A.M. – 12:00 P.M.
Purdue **G**raduate **S**tudent **C**enter
504 Northwestern Ave. ~ Room 105A/B

Deep Learning Training and Inference Optimization Towards Speed and Scale

Abstract: The application of deep learning models significantly improves many services and products. However, it is challenging to provide efficient computation and memory capabilities for both DNN workload training and inference, given that the model size and complexities keep increasing. From the training aspect, it is too slow to train high-quality models on massive data, and large-scale model training often requires complex refactoring of models and access to prohibitively expensive GPU clusters, which are not always accessible to many practitioners. On the serving side, many DL models suffer from long inference latency and high costs, preventing their deployment in production. In this talk, I will introduce my experience and learning from designing and implementing optimizations for both DNN training and serving at large scale with remarkable compute and memory efficiency improvement and infrastructure cost reduction.

Bio: Dr. Minjia Zhang is working as a researcher at Microsoft. His research interest lies in AI systems, algorithms, and their applications in large-scale natural language processing and information retrieval. In particular, he focuses on building efficient and effective DL training libraries such as DeepSpeed and ultra-fast and high throughput DL inference acceleration libraries such as DeepCPU. Several of his research results have been transferred to Microsoft systems and products, such as Bing, Ads, Azure SQL, Windows, leading to significant latency and capacity improvement.