The Department of Electrical and Computer Engineering Presents

A DISTINGUISHED SEMINAR

LEANA GOLUBCHIK

DECONSTRUCTING DISTRIBUTED DEEP LEARNING

ABSTRACT

Deep learning has made substantial strides in computer vision, speech recognition, natural language processing, and other applications. New training techniques, larger datasets, increased computing power, and easy-to-use machine learning frameworks (such as TensorFlow, PyTorch or Caffe) all contribute to this success. An important missing piece is that deep learning frameworks do not assist the user with provisioning and sharing cloud resources, or with the integration of DNN training workloads into existing datacenters. Most users need to try different configurations of a job (such as number of server/worker nodes, mini-batch size, network capacity) to determine the resulting training performance (throughput measured as examples/second and training accuracy). When resources must be shared among hundreds of jobs, this approach quickly becomes infeasible. At a larger scale, when multiple datacenters need to manage deep learning workloads, different degrees of affinity for their resources create economic incentives to collaborate, as in cloud federations. In this talk, we present recent models to predict performance metrics (such as training throughput) and scheduling algorithms that use these metrics to guide resource allocation. We also outline the economic incentives for resource sharing for such workloads, and future research goals to broaden the population of users capable of discovering deep learning models and applying them to novel applications.

BIOGRAPHY

Leana Golubchik is the Stephen and Etta Varra Professor of Computer Science and Electrical Engineering at USC. She also serves as the Director of the Women in Science and Engineering (WiSE) program. Prior to that, she was on the faculty at the University of Maryland and Columbia University. Leana received her Ph.D. from UCLA. Her research interests are broadly in the design and evaluation of large scale distributed systems, including hybrid clouds and data centers and their applications in data analytics, machine learning, and more recently privacy. Leana received several awards, including the IBM Faculty Award, the NSF CAREER Award, the Okawa Foundation Award, the WTS-LA Diversity Leadership Award, the USC Remarkable Women Award, and the USC Mellon Culture of Mentoring Award. She is on the the Editorial Boards of the ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS) and the Performance Evaluation journal as well as a member of the IFIP WG 7.3 (elected in 2000).

April 8, 2019

WCH 205/206

11:10 a.m. - 12:00 p.m.