ECE Colloquium

ITE 336

11:15 am Friday March 1, 2024

(Refreshments in ITE 301 @ 11 am)

Computationally Efficient Deep Learning with Theoretical Generalization Guarantees

Dr. Meng Wang, Rensselaer Polytechnic Institute

Deep learning has demonstrated extraordinary empirical success in various applications. Despite the great promise, DL has significant computation requirements. Moreover, DL lacks interpretability and performance guarantees and often acts as the “black box” due to its extreme complexity. This talk introduces our ongoing efforts to reduce the computation cost of deep learning while maintaining its ability to generalize on unseen data. Our first approach is network pruning, which removes part of the neuron weights in a complex model to reduce the inference cost. We illustrate how the degree of network sparsity impacts sample complexity and convergence rates quantitatively. Furthermore, we establish that typical magnitude-based pruning methods can effectively reduce model size without compromising generalization performance across a broad range of scenarios. Our second approach is the Mixture of Experts (MoE) architecture, which routes each input to a few subnetworks (termed as experts) rather than processing it through the entire network to reduce computation. We explore the application of patch-level MoEs in vision and language tasks and offer the first theoretical guarantee of their generalization performance. We demonstrate that patch-level MoEs simultaneously reduce sample complexity, computational complexity, and model complexity while maintaining equivalent generalization accuracy.

Meng Wang is an Associate Professor in the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute. She received B.S. and M.S. degrees from Tsinghua University, China, in 2005 and 2007, respectively, and a Ph.D. degree from Cornell University, Ithaca, NY, USA, in 2012. Prior to joining RPI in December 2012, she was a postdoc research scholar at Duke University. Her research areas include machine learning and data analytics, energy systems, signal processing, and optimization. She is a recipient of the Young Investigator Program (YIP) Awards from the Air Force Office of Scientific Research (AFOSR) in 2019 and the Army Research Office (ARO) in 2017. At Rensselaer, she received the James M. Tien '66 Early Career Award and Grant for Faculty in 2022 and the School of Engineering Research Excellence Award in 2018. She has been an Associate Editor of IEEE Transactions on Smart Grids since 2020 and was a guest editor of the IEEE Journal of Selected Topics in Signal Processing Special Issue on Signal and Information Processing for Critical Infrastructures in 2018.