NJIT

Department of Computer Science

Resource-efficient Deep Learning: Democratizing AI at Scale

Dongkuan Xu
Pennsylvania State University

Hosted by Przemyslaw Musialski

DATE: Monday, February 21, 2022
COFFEE: 2:15 PM – 2:30 PM
TIME: 2:30 PM – 3:30 PM
LOCATION: GITC 4402 (4th Floor Seminar Lecture Hall)
WEBEX LINK: https://njit.webex.com/njit/j.php?MTID=m4593df338343f34f6a3f1dc69534a1f5

http://cs.njit.edu/seminars

Abstract: The phenomenal success of deep learning in the past decade has been mostly driven by the construction of increasingly large deep neural network models. These models usually impose an ideal assumption that there are sufficient resources, including large-scale parameters, sufficient data, and massive computation, for the optimization. However, this assumption usually fails in real-world scenarios. For example, computer memory may be limited as in edge devices, large-scale data are difficult to obtain due to expensive costs and privacy constraints, and computational power is constrained as in most university labs. As a result, these resource discrepancy issues have hindered the democratization of deep learning techniques in many AI applications, and the development of efficient deep learning methods that can adapt to different resource constraints is of great importance.

In this talk, I will present my recent research contributions centered around resource-efficient deep learning to free AI from the parameter-data-computation hungry beast. First, I will introduce the significance of my contribution on neural network pruning, which improves the parameter efficiency of large-scale pre-trained language models in the inference phase, resulting in pruned models with an order-of-magnitude fewer parameters than the original model while achieving the same or better prediction accuracy. Then, I will talk about my task-agnostic neural architecture search framework to reduce the computational cost in the training phase for finding the best pruned models, which is complementary to improving the parameter efficiency in the inference phase. Finally, I will conclude my presentation with a brief overview of my ongoing and future work as part of a broader research agenda of new and related problems and potential collaborations in the next few years.

Bio: Dongkuan (DK) Xu is a Ph.D. student at Penn State, advised by Prof. Xiang Zhang. His research interest is resource-efficient deep learning for AI at scale, focusing on how to improve the efficiency of deep learning systems to achieve Pareto optimality between resources (e.g., parameters, data, computation) and performance (e.g., inference, training). DK has published more than 25 papers in top conferences and journals, including NeurIPS, AAAI, ACL, NAACL, and IJCAI. He has served as a (senior) PC member or regular reviewer for over 28 major conferences and 14 journals, and has worked as an instructor or teaching assistant for 8 courses. DK also has extensive research experience in industry. He has interned at Microsoft Research

Redmond, Moffett AI, and NEC labs America, and holds 8 US patents/applications. DK's long-term research goal is to democratize AI to serve a broader range of populations and domains.