香港城市大學
City University of Hong Kong

電腦科學系
Department of
Computer Science

COMPUTER SCIENCE COLLOQUIUM

Optimizing Device Placement in Machine Learning Workloads using Deep Reinforcement Learning

SPEAKER **Prof Baochun LI**

Professor
Department of Electrical and
Computer Engineering
University of Toronto
Canada

DATE 17 December 2018 (Monday)
TIME 10:30 am - 11:30 am
VENUE CSE Conference Room , B6605
6th Floor, Blue Zone
Yeung Kin Man Academic Building
City University of Hong Kong
83 Tat Chee Avenue
Kowloon Tong

ABSTRACT

Training deep neural networks requires an exorbitant amount of computation resources, including a heterogeneous mix of GPU and CPU devices. It is critical to place operations in a neural network on these devices in an optimal way, so that the training process can complete within the shortest amount of time. The state-of-the-art in the literature uses a deep reinforcement learning method based on policy gradients to solve this problem, but we believe that there remains ample room for further improvements. In this talk, I will present our recent work published in ICML 2018 and NeurIPS 2018 that uses proximal policy optimization (PPO) and cross-entropy minimization to achieve significantly better performance than the state-of-the-art. Our experiments with several popular neural network training benchmarks have demonstrated clear evidence of superior performance: with the same amount of learning time, our algorithm leads to placements that have training times up to 60% shorter. This talk will be self-contained, starting with a quick tutorial on basic ideas in reinforcement learning algorithms.

BIOGRAPHY

Baochun Li received his B.Engr. degree from the Department of Computer Science and Technology, Tsinghua University, China, in 1995 and his M.S. and Ph.D. degrees from the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, in 1997 and 2000. Since 2000, he has been with the Department of Electrical and Computer Engineering at the University of Toronto, where he is currently a Professor. He holds the Bell Canada Endowed Chair in Computer Engineering since August 2005. His research interests include cloud computing, distributed systems, datacenter networking, and wireless systems.

Prof Li has co-authored more than 360 research papers, with a total of over 17000 citations, an H-index of 75 and an i10-index of 233, according to Google Scholar Citations. He was the recipient of the IEEE Communications Society Leonard G. Abraham Award in the Field of Communications Systems in 2000. In 2009, he was a recipient of the Multimedia Communications Best Paper Award from the IEEE Communications Society, and a recipient of the University of Toronto McLean Award. He is a member of ACM and a Fellow of IEEE.

All are welcome!

In case of questions, please contact Dr Cong Wang at Tel: 3442 2010, E-mail: congwang@cityu.edu.hk, or visit the CS Departmental Seminar Web at http://www.cs.cityu.edu.hk/.