COMPUTER SCIENCE COLLOQUIUM

From 7,000X Model Compression to 100X Acceleration – Achieving Real-Time Execution of ALL DNNs on Mobile Devices

SPEAKER
Dr. WANG Yanzhi
Assistant Professor
Department of Electrical and Computer Engineering
Northeastern University
USA

DATE
10 October 2019 (Thursday)

TIME
10:00 am - 11:00 am

VENUE
CS Seminar Room, Y6405, LT-2, 4th Floor, 6th Floor, Mr and Mrs Sze Chi Ching Lecture Theatre, Yeung Kin Man Academic Building, City University of Hong Kong
83 Tat Chee Avenue
Kowloon Tong

ABSTRACT
This presentation focuses on two recent contributions on model compression and acceleration of deep neural networks (DNNs). The first is a systematic, unified DNN model compression framework based on the powerful optimization tool ADMM (Alternating Direction Methods of Multipliers), which applies to non-structured and various types of structured weight pruning as well as weight quantization technique of DNNs. It achieves unprecedented model compression rates on representative DNNs, consistently outperforming competing methods. When weight pruning and quantization are combined, we achieve up to 6,635X weight storage reduction without accuracy loss, which is two orders of magnitude higher than prior methods. Our most recent results conducted a comprehensive comparison between non-structured and structured weight pruning with quantization in place, and suggest that non-structured weight pruning is not desirable at any hardware platform.

However, using mobile devices as an example, we show that existing model compression techniques, even assisted by ADMM, are still difficult to translate into notable acceleration or real-time execution of DNNs. Therefore, we need to go beyond the existing model compression schemes, and develop novel schemes that are desirable for both algorithm and hardware. Compilers will act as the bridge between algorithm and hardware, maximizing parallelism and hardware performance. We develop a combination of pattern pruning and connectivity pruning, which is desirable at all of theory, algorithm, compiler, and hardware levels. We achieve 18.9ms inference time of large-scale DNN VGG-16 on smartphone without accuracy loss, which is 55X faster than TensorFlow-Lite. We can potentially enable 100X faster and real-time execution of all DNNs using the proposed framework.

BIOGRAPHY
Yanzhi Wang is currently an assistant professor in the Department of Electrical and Computer Engineering at Northeastern University. He has received his Ph.D. Degree in Computer Engineering from University of Southern California (USC) in 2014, and his B.S. Degree with Distinction in Electronic Engineering from Tsinghua University in 2009.

Dr. Wang’s current research interests mainly focus on DNN model compression and energy-efficient implementation (on various platforms). His research maintains the highest model compression rates on representative DNNs since 09/2018. His work on AQFP superconducting based DNN acceleration is by far the highest energy efficiency among all hardware devices. His work has been published broadly in top conference and journal venues (e.g., ASPLOS, ISCA, MICRO, HPCA, ISSCC, AAAI, ICML, CVPR, ICLR, IJCAI, ECCV, ICDM, ACM MM, DAC, ICCAD, FPGA, LCTES, CCS, VLDB, ICDCS, TComputer, TCAD, JSAC, TNNLS, Nature SP, etc.), and has been cited around 5,000 times. He has received four Best Paper Awards, has another eight Best Paper Nominations and three Popular Paper Awards.

All are welcome!

In case of questions, please contact Dr Chun Jason XUE at Tel: 3442 9815, E-mail: jasonxue@cityu.edu.hk, or visit the CS Departmental Seminar Web at http://www.cs.cityu.edu.hk/.