北京大学高能效计算与应用中心学术报告

Invited Talk, Center for Energy-Efficient Computing and Applications

FROM 7,000X MODEL COMPRESSION TO 100X ACCELERATION

ACHIEVING REAL-TIME EXECUTION OF ALL DNNS ON MOBILE DEVICES

Prof. Yanzhi Wang

Department of Electrical and Computer Engineering

Northeastern University

2019年11月13日 星期三 13:30pm

理科二号楼2135会议室

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, which applies to non-structured and various types of structured weight pruning as well as weight quantization. It achieves unprecedented model compression rates on representative DNNs. When weight pruning and quantization are combined, we achieve up to 6,635X weight storage reduction without accuracy loss. 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. We develop a combination of pattern 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, 55X faster than TensorFlow-Lite.

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.