YNU 横浜国立大学 YOKOHAMA National University

Institute of Advanced Sciences Yokohama National University

Institute of Advanced Sciences
先端科学高等研究院

2019-2nd IAS-YNU Seminar of Research Unit for Extremely Energy-Efficient Processors
超省エネルギープロセッサ研究ユニット 2019年度第2回 IAS-YNUセミナー

Title of Talk: **5,000X model compression in DNNs; But, is it truly desirable?**

Speaker: **Prof. Yanzhi Wang**
Depart. of Electrical and Computer Eng.
Northeastern University, USA

Date: 13:30 – 14:20, June 14 (Fri), 2019
日程: 令和元年6月14日（金）13:30～14:20

Place: Electrical & Computer Eng. Bldg. 4F (Seminar room I)
場所: 電子情報工学棟４階（演習室 I）

Abstract:

Hardware implementation of deep neural networks (DNNs) with emphasis on performance and energy efficiency has been the focus of extensive ongoing investigations. When large DNNs are mapped to hardware as an inference engine, the resulting hardware suffers from significant performance and energy overheads. To overcome this hurdle, we develop ADMM-NN, an algorithm-hardware co-optimization framework for greatly reducing DNN computation and storage requirements by incorporating Alternating Direction Method of Multipliers (ADMM) and utilizing all redundancy sources in DNN. Our preliminary results show that ADMM-NN can achieve the highest degree of model compression on representative DNNs. For example, we can achieve 348X, 63X, 34X, and 17X weight reduction on LeNet-5, AlexNet, VGGNet, and ResNet-50, respectively, with (almost) no accuracy loss. We achieve a maximum of 4,438X weight data storage reduction when combining weight pruning and weight quantization, while maintaining accuracy. However, a second problem arises. The needs for index storage is even higher compared with weight value storage, especially when weight quantization is in place. Then the question is: is it possible that incorporating "structures" in weight pruning results in even loss storage compared with the general, non-structured pruning? If the answer is yes, then whether non-structured pruning is still a viable approach at all? Through extensive investigation, we found that the answer is yes for most cases under the same accuracy. As a result, we recommend not to continue DNN inference engines based on non-structured sparsity.

問合せ先：電情　吉川