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

EFFICIENT METHODS AND HARDWARE FOR DEEP LEARNING

& MIT EECS Ph.D. Recruitment Seminar

Dr. Song Han

Electrical Engineering & Computer Science Department

MIT

2017年9月13日 星期三 02:00pm

第三教学楼 401教室

ABSTRACT: Deep learning has spawned a wide range of AI applications that are changing our lives. However, deep neural networks are both computationally and memory intensive. Thus they are power hungry when deployed on embedded systems and data centers with a limited power budget. To address this problem, I will present an algorithm and hardware co-design methodology for improving the efficiency of deep learning.

I will first introduce "Deep Compression", which can compress deep neural network models by 10–49× without loss of prediction accuracy for a broad range of CNN, RNN, and LSTMs. The compression reduces both computation and storage. Next, by changing the hardware architecture and efficiently implementing Deep Compression, I will introduce EIE, the Efficient Inference Engine, which can perform decompression and inference simultaneously, saving a significant amount of memory bandwidth. By taking advantage of the compressed model and being able to deal with an irregular computation pattern efficiently, EIE achieves 13× speedup and 3000× better energy efficiency over GPU. Finally, I will revisit the inefficiencies in current learning algorithms, present DSD training, and discuss the challenges and future work in efficient deep learning.

MIT EECS department ranks top in the US. At the end of the seminar, I will introduce the Ph.D. recruitment in the area of artificial intelligence and computer architecture at MIT. I have multiple openings for both Ph.D. students and summer interns.

BIOGRAPHY: Song Han graduated from Stanford University advised by Prof. Bill Dally. He will join MIT EECS as assistant professor in July 2018. His research focuses on energy-efficient deep learning, at the intersection between machine learning and computer architecture. He proposed the Deep Compression algorithm, which can compress neural networks by 18-49× while fully preserving prediction accuracy. He designed the first hardware accelerator that can perform inference directly on a compressed sparse model, which results in significant speedup and energy saving. His work has been featured by O'Reilly, TechEmergence, TheNextPlatform, and Embedded Vision, and it has impacted the industry. He led research efforts in model compression and hardware acceleration that won the Best Paper Award at ICLR'16 and the Best Paper Award at FPGA'17. Before joining Stanford, Song graduated from Tsinghua University.