Hardware-Centric AutoML: Design Automation for Efficient Deep Learning Computing

**SPEAKER**  
Dr Song Han  
Assistant Professor  
Department of Electrical Engineering and Computer Science  
Massachusetts Institute of Technology  
USA

**DATE**  
31 January 2019 (Thursday)

**TIME**  
10:00 am - 11:00 am

**VENUE**  
CS Seminar Room, Y6405  
6th Floor, Yellow Zone  
Yeung Kin Man Academic Building  
City University of Hong Kong  
83 Tat Chee Avenue  
Kowloon Tong

**ABSTRACT**

In the post-ImageNet era, researchers are solving more complicated AI problems using larger data sets which drives the demand for more computation. However, Moore’s Law is slowing down. The mismatch between supply and demand for computation highlights the need for co-designing efficient machine learning algorithms and domain-specific hardware architectures. We introduce our recent work using machine learning to optimize the machine learning system (Hardware-centric AutoML): learning the optimal pruning strategy (AMC) and quantization strategy (HAQ) on the target hardware; learning the optimal neural network architecture that is specialized for a target hardware architecture (ProxylessNAS); learning to optimize analog circuit parameters, rather than relying on experienced analog engineers to tune those transistors (L2DC). For hardware-friendly machine learning algorithms, I’ll introduce the temporal shift module (TSM) for efficient video understanding, that offers 8x lower latency, 12x higher throughput than 3D convolution-based methods. I’ll describe efficient deep learning accelerators that can take advantage of these efficient algorithms, including both FPGA and ASIC designs for emerging deep learning architectures. I’ll conclude the talk by giving an outlook of the design automation for efficient deep learning computing.

**BIOGRAPHY**

Dr. Song Han is an assistant professor in the EECS Department of Massachusetts Institute of Technology (MIT) and PI for HAN Lab: Hardware, AI and Neural-nets. Dr. Han’s research focuses on energy-efficient deep learning and domain-specific architectures. He proposed “Deep Compression” that widely impacted the industry. He was the co-founder and chief scientist of DeepPhi Tech based on his PhD thesis. Prior to joining MIT, Song Han graduated from Stanford University.

All are welcome!

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