Hardware-aware Neural Architecture Search and Compression for Efficient Deep Learning
Song Han, Assistant Professor, MIT EECS

Efficient deep learning computing requires algorithm and hardware co-design to enable specialization. However, the extra degree of freedom creates a much larger design space. Human engineers can hardly exhaust the design space by heuristics. We propose AutoML techniques to architect efficient neural networks. We investigate automatically designing small and fast models (ProxylessNAS), auto channel pruning (AMC), and auto mixed-precision quantization (HAQ). We demonstrate such learning-based, automated design achieves superior performance and efficiency than rule-based human design. Moreover, we shorten the design cycle by 200× than previous work, so that we can afford to design specialized neural network models for different hardware platforms. Finally, we accelerate computation-intensive AI applications including TSM for efficient video recognition and PVCNN for efficient 3D recognition on point clouds.

Bio: Song Han is an assistant professor at MIT EECS. Dr. Han received the Ph.D. degree in Electrical Engineering from Stanford advised by Prof. Bill Dally. Dr. Han's research focuses on efficient deep learning computing. He proposed “Deep Compression” and “ EIE Accelerator" that impacted the industry. His work received the best paper award in ICLR'16 and FPGA’17. He was the co-founder and chief scientist of DeePhi Tech which was acquired by Xilinx.