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 LEARNING FOR THE EDGE

Bichen Wu, PhD Candidate

EECS, UC Berkeley

Host：孙广宇 长聘副教授

2019年1月4日 星期五 10:00am

理科五号楼 410会议室

ABSTRACT: The success of deep neural networks is attributed to three factors: stronger computing capacity, more complex neural networks, and more data. These factors, however, are usually not available when we apply DNNs to edge applications such as autonomous driving, augmented and virtual reality (AR/VR), internet-of-things (IoT), and so on. Training DNNs requires a large amount of data, which can be difficult to obtain. Edge devices such as mobile phones or IoT devices have limited computing capacity, which requires specialized and efficient DNNs. However, due to the diversity and complexity of hardware devices, the enormous design space of DNNs, and prohibitive training costs, designing efficient DNNs for target devices is a challenging task. In this talk, we introduce our recent works addressing these problems. First, we introduce SqueezeSegV2, an efficient DNN for LiDAR point cloud segmentation for autonomous driving. LiDAR point cloud data are extremely difficult to annotate. We bypass this by leveraging simulated data to train the network and adapting it to achieve a performance comparable to training on real data. Second, we introduce Synetgy, a DNN-model & hardware co-designed FPGA accelerator that achieves 16.9x speedup over the previous state-of-the-art. Lastly, we introduce DNAS, a differentiable neural architecture search framework for automatic DNN design. With a small computational cost (8GPUs for 1 day), DNAS discovers a family of DNNs called FBNet that outperform previous state-of-the-art models designed manually and automatically. For different target devices, DNAS automatically adapt DNN architectures accordingly to optimize for latency while maintaining accuracy.

BIOGRAPHY: Bichen Wu is a PhD candidate at EECS, UC Berkeley. He works with Prof. Kurt Keutzer, and he is affiliated with Berkeley AI Research (BAIR) and Berkeley Deep Drive (BDD). His research focus is on efficient deep learning, computer vision, and autonomous driving.