Once for All: Train One Network and Specialize it for Efficient Deployment

Song Han, Assistant Professor, MIT EECS

We address the challenging problem of efficient deep learning model deployment, where the goal is to design neural network architectures that can fit different hardware platform constraints. Most of the traditional approaches either manually design or use Neural Architecture Search (NAS) to find a specialized neural network and train it from scratch for each case, which is computationally expensive and unscalable. Our key idea is to decouple model training from architecture search to save the cost. To this end, we propose to train a once-for-all network (OFA) that supports diverse architectural settings (depth, width, kernel size, and resolution). Given a deployment scenario, we can then quickly get a specialized sub-network by selecting from the OFA network without additional training. To prevent interference between many sub-networks during training, we also propose a novel progressive shrinking algorithm, which can train a surprisingly large number of sub-networks (> 10^{19}) simultaneously, while maintaining the same accuracy as independently trained networks. Extensive experiments on various hardware platforms (CPU, GPU, mCPU, mGPU, FPGA accelerator) show that OFA consistently achieves the same level (or better) ImageNet accuracy than SOTA NAS methods while reducing orders of magnitude GPU hours and CO_2 emission than NAS. In particular, OFA requires 16x fewer GPU hours than ProxylessNAS, 19x fewer GPU hours than FBNet and 1,300x fewer GPU hours than MnasNet under 40 deployment scenarios.