Compressed CNN Training with FPGA-based Accelerator

Kaiyuan Guo, Shuang Liang, Jincheng Yu, Xuefei Ning, Wenshuo Li, Yu Wang, Huazhong Yang, *Tsinghua University*
Contact: gky15@mails.tsinghua.edu.cn

Training convolutional neural network (CNN) usually requires large amount of computation resource, time and power. Researchers and cloud service providers in this region needs fast and efficient training system. GPU is currently the best candidate for CNN training. But FPGAs have already shown good performance and energy efficiency as CNN inference accelerators. In this work, we design a compressed training process together with an FPGA-based accelerator for energy efficient CNN training. We adopt two of the widely used model compression methods, quantization and pruning, to accelerate CNN training process.

The difference between inference and training brought challenges to apply the two methods in training. First, training requires higher data precision. We use the gradient accumulation buffer to achieve low operation complexity while keeping gradient descent precision. Second, sparse network results in different types of functions in forward and back-propagation phases. We design a novel architecture to utilize both inference and back-propagation sparsity. Experimental results show that the proposed training process achieves similar accuracy compared with traditional training process with floating point data. The proposed accelerator achieves 641GOP/s equivalent performance and 2.86x better energy efficiency compared with GPU.

Keywords: FPGA; Convolutional Neural Network; Training

DOI: https://doi.org/10.1145/3289602.3293977