Fast Machine Learning for Science
Imperial College London
Real-time and accelerated ML for fundamental sciences
25-28 September 2023

Contribution ID: **44**

Type: **Standard Talk**

Post-training ReLU Sparsification for Faster CNN Inference on FPGA Streaming Accelerators

Wednesday 27 September 2023 16:45 (15 minutes)

Convolutional Neural Networks (CNNs) have been applied to a wide range of applications in high energy physics including jet tagging and calorimetry. Due to their computational intensity, a large amount of work has been done to accelerate CNNs in hardware, with FPGA devices serving as a high-performance and energy-efficient platform of choice. As opposed to a dense computation where every single multiplication in a convolution is performed, there is a large proportion of zero values in the activation maps due to the ReLU activation layers. Recent work has explored threshold-based sparsification of CNNs by retraining with a parameterized activation function, termed FATReLU which zero-outs all values less than a positive threshold for more sparsity. In this work, we explore the use of ReLU threshold-based sparsification without retraining instead as a time-efficient method and present a sparse accelerator toolchain based on fpgaConvNet.

At a macro level, the accelerator partitions a CNN model and for each partition, it stores all weights in on-chip memory and executes model inference in a layer-wise pipeline. At a micro level, it dynamically skips zero-valued multiplications in hardware via a non-zero check and a crossbar switch to speed-up computation at runtime. We model the performance benefits and measure accuracy loss of ReLU-based sparsification on hardware implementations of ResNet-18 and ResNet-50 on a Xilinx Alveo U250 board. In doing so, we demonstrate the accuracy benefit of latency-aware thresholding, where ReLU thresholds are iteratively increased to boost the sparsity of each partition's slowest node. We measure the performance gains of our method on the baseline-hardware as well as on optimised-hardware, which refers to the new design obtained after performing a design space exploration for the boosted sparsity.

Compared to existing sparse-accelerated designs with the same resources, we observe upto 16% and 29% increase in throughput for the baseline-hardware and optimised-hardware designs, respectively for < 1% loss in Top-1 accuracy on CIFAR-10 image classification without any model retraining. Using latency-aware thresholding, we observe upto 23% and 36% increase in throughput for baseline-hardware and optimised-hardware designs, respectively for the same accuracy. Our work demonstrates that post-training ReLU-based sparsification provides a cheap and useful trade-off between performance and accuracy in sparse CNN inference. A summary of the results can be viewed here: https://drive.google.com/file/d/1nrEoCvD09nku-SVYVPAPlg1OWi_kUsbp/view?usp=sharing

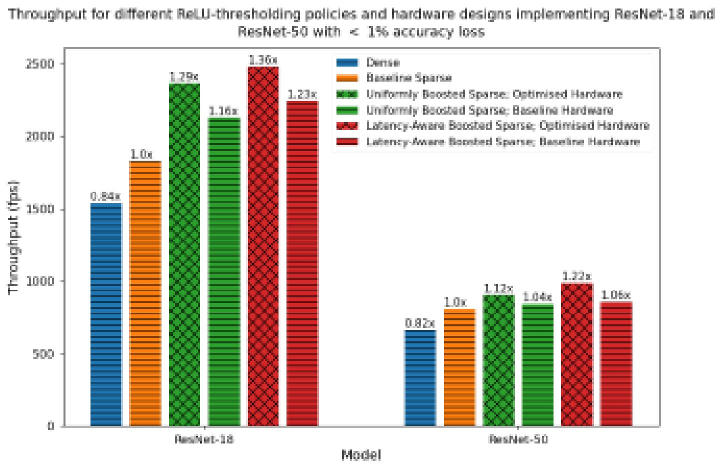

Figure 1: Summary of Throughput Results for ResNet-18 and ResNet-50 Implementations on Xilinx Alveo U250 for CIFAR-10 Image Classification

Authors: Mr AGRAWAL, Krish (Imperial College London); Mr YU, Zhewen (Imperial College London); Mr MONTGOMERIE-CORCORAN, Alexander (Imperial College London); Dr BOUGANIS, Christos-Savvas (Imperial College London)

Presenters: Mr AGRAWAL, Krish (Imperial College London); Mr YU, Zhewen (Imperial College London); Mr MONTGOMERIE-CORCORAN, Alexander (Imperial College London)

Session Classification: Contributed Talks

Track Classification: Contributed Talks