ABOUT | CONTRIBUTORS | CONTACT | THE REGISTER

THENEXTPLATFORM

HOME COMPUTE STORE CONNECT CONTROL CODE ANALYZE HPC ENTERPRISE HYPERSCALE C

FPGAS FOCAL POINT FOR EFFICIENT NEURAL NETWORK INFERENCE

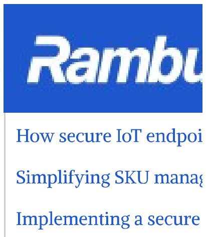
January 26, 2017 Nicole Hemsoth

Over the last couple of years, we have focused extensively on the hardware required for training deep neural networks and other machine learning algorithms. Focal points have included the use of general purpose and specialized CPUs, GPUs, custom ASICs, and more recently, FPGAs.

As the battle to match the correct hardware devices for these training workloads continues, another has flared up on the deep learning inference side. Training neural networks has its own challenges that can be met with accelerators, but for inference, the efficiency, performance, and accuracy need to be in balance.

One developing area in inference is in the use of binarized neural networks (BNNs)—an approach that has implications on the low-precision side for both hardware and software. In a recent analysis of the use of BNNs to accelerate inference, deep learning chip maker, Nervana (now part of Intel), described the benefits with an emphasis on efficiency:

> "BNNs use binary weights and activations for all computations. Floating point arithmetic underlies all computations in deep learning, including computing gradients, applying parameter updates, and calculating activations. These 32-bit floating point multiplications, however, are very expensive. In BNNs, floating point multiplications are supplanted with bitwise XNOR and



> left and right bit shifts…This is extremely attractive from a hardware perspective: binary operations can be implemented computationally efficiently at a low power cost.”

There are multiple ways this could be implemented from a hardware perspective, but low-precision capabilities are critical. Nervana’s own technology (which we described here) is one option, the low-precision capabilities in the newest Pascal GPUs is another, but FPGAs hold promise here as well. A new investigation into how FPGAs might fit the BNN bill has emerged from FPGA maker Xilinx researchers, along with research partners at the University of Sydney and the Norwegian University of Science and Technology.

“Binarized neural networks are gaining interest in the deep learning community due to their significantly lower computational and memory cost,” the FPGA researchers explain. “They are particularly well suited to reconfigurable logic devices, which contain an abundance of fine-grained compute resources and can result in smaller, lower power implementations, or conversely, in higher classification rates.” The team’s results, which are based on the Xilinx ADM-PCIE-8K5 show “very high image classification rates, minimal latency, and very high power efficiency can be achieved by mapping BNNs to FPGAs, even though improvements still may be made.”

> “FPGAs have a much higher theoretical peak performance for binary operations compared to floating point, while the small memory footprint removes the off-chip memory bottleneck by keeping parameters on chip, even for large networks… BNNs are particularly appealing since they can be implemented almost entirely with binary operations, with the potential to attain performance in the teraops per second range on FPGAs.”

At the root of this work is a framework for efficient, scalable BNN inference on an FPGA the Xilinx researchers developed, called FINN. As they described in their initial paper, “FINN-generated accelerators can perform millions of classifications per second with sub-microsecond latency, thereby making them ideal for supporting real-time embedded applications, such as augmented reality, autonomous driving, and robotics.” The work done here fed into the more recent BNN research, which aims a more powerful FPGA at the BNN problem and tests it at greater scale and network complexity. While they were able to demonstrate positive results with the larger scale FINN FPGA effort for use in a datacenter environment (on deep learning inference applications), there are still some kinks, which once worked out, will lead to even greater inference efficiency.

While this research is promising, there are still areas where researchers see a need for further development. BRAM utilization across their experiments is not high, which will become an ever-larger problem as the scale of the problem size goes up. Further, as the team notes, “The downside to the high performance characteristics of BNNs is a small drop in accuracy, in comparison to floating point networks. Improving the accuracy for reduced precision CNNs is an active research area in the machine learning community and first evidence shows that accuracy can be improved by increasing network sizes.”

For those interested, the benchmarks for efficiency and performance are offered in detail, along with details about how the padding and network sizes have been tweaked to work well with datacenter-scale deep learning inference problems.

For background on binarized neural networks, there is an excellent presentation from NIPS 2016 from some of the leading researchers in this area here.