**Energy-efficient Neural Network Accelerator Based on Outlier-aware Low-precision Computation**

**Authors**

1. **Eunhyeok Park**  Seoul National University
2. **Dongyoung Kim** Seoul National University
3. **Sungjoo Yoo**  Seoul National University

**Keywords**

1. Convolutional neural networks (CNN)
2. Inference acceleration
3. Outlier-aware quantization
4. Algorithm-Architecture co-design

**Summary**

*Challenge*

Owning to the presence of large values, which we call outliers, conventional methods of quantization fail to achieve significantly low precision, e.g., four bits for very deep neural networks, such as ResNet-101. However, a novel quantization method called *outlier-aware quantization*, which divides the distribution of data (weights and activations) into two regions, of low and high precision. This quantization method enables low precision, e.g. four bits, for very deep models, such as ResNet-101 and DensNet-121, with a very small ratio (3%) of outliers at a negligible (<1%) loss of accuracy.

*Contributions*

This paper proposed accelerator, called the outlier-aware accelerator (OLAccel), performs dense and reduced precision computations on a majority of data in the low-precision region while performing sparse and high-precision computation on outliers, which achieves a significant reduction in energy consumption by reducing the amount of memory access, and by using smaller units of computation. The main contributions of this paper are as follows:

1. This paper proposes an accelerator called OLAccel that implements 4-bit computations on very deep CNNs which differently handling outlier activations and weights to improve computational efficiency.
2. OLAccel could perform sparse high-percision computation for outlier activations in parallel with dense low-precision computation for a majority of activations, which could reduce latency. This is achieved through equipping SIMD lanes with an outlier MAC unit.
3. This paper designs OLAccel in Verilog and compare it with prevalent accelerators, Eyeriss and ZeNA at reduced precision.

*Experiments and Results*

This paper implements outlier-aware quantization in the PyTorch and Caffe, and it calculate average energy consumption and execution cycles by running neural networks on hardware accelerator models.

For the baseline and the paper’s architecture, this paper developed cycle-accurate models, and does a Verilog design and synthesized them using Design Compiler with a commercial 65-nm LP library at 250MHz and 1.0V. This paper used CACTI to estimate SRAM area/power/latency and the DRAM power model from Micron.

This paper compare 4-bit OLAccel with the following baselines:

1. 16-bit and 8-bit Eyeriss (denoted by Eyeriss16 and Eyeriss8, respectively), which, for zero input, do not reduce the number of execution cycles but clock-gate computations.
2. 16-bit and 8-bit ZeNAs that reduce the number of execution cycles by skipping computations with zero input weights and activations.

The results show that the OLAccel can reduce by 43.5% (27.0%), 56.7% (36.3%), and 62.2% (49.5%) energy consumption for AlexNet, VGG-16, and ResNet-18, respectively, compared with a 16-bit (8-bit) state-of-the-art zero-aware accelerator. The energy gain mostly comes from the memory components, the DRAM, and on-chip memory due to reduced precision.

*Comments*

There is not much to say about this paper. This paper has realized a new CNN accelerator to support a novel quantization method and could greatly save energy consumption, there it can be used in practical application scenarios. However, this architecture need to be specifically designed and implemented, therefore losing generality compared with 8-bit or other accelerators.