Precision Reconfigurable Digital Compute-In-Memory for Embedded Neural Network Processing

Authors

Bongjin Kim (Assistant Professor) [bjkim@ntu.edu.sg], Hyunjoon Kim (Ph.D Student) [kimh0003@e.ntu.edu.sg], Nanyang Technological University, Singapore

Category

Hardware & Systems

Problem

High energy and area efficiency, robustness, reconfigurability, and scalability are key specifications for machine learning (ML) hardware accelerator design for edge computing applications. Embedding ML cores into edge computing devices present great challenges in dealing with energy/area constrained environments and performance reliability during deployment. Engineers and researchers struggled to deliver targeted specs through both algorithmic improvements as well as hardware micro architecture advancements.

One of the key findings that we focused was network quantization. Various quantization strategies have been explored in order to achieve higher efficiency while minimizing the prediction accuracy loss in image classification tasks (i.e. ImageNet). Our target was to realize per-layer quantization through reconfigurable hardware which reduces the computation complexity while having minimal loss in prediction accuracy. We attempt to address the key target specifications mentioned above by utilizing different circuit design techniques, improved algorithmic and micro architecture efforts.

Novelty

We propose a column-MAC structure dot-product compute engine that can be reconfigured from 1-to-16bit. The design utilizes unit bit-processing block (a bitcell), which can be stacked or decomposed to process 1-to-16bit weights and inputs. As an attempt to avoid increasing hardware and computation complexity of dot product in higher bit-precision numbers, we implemented bit-serial computation scheme with parallelism to mitigate the trade-off in latency and throughput. Conventional digital accelerators' concern of having large power consumption from off-chip DRAM access was addressed by implementing compute-in memory architecture. In order to further improve the performance, we also introduced sparse pipelining and unique combination of number representations for weights and inputs.

The proposed bitcell array enables scalable, reconfigurable and energy-efficient compute-in-memory operation for neural network applications using custom digital circuits. The proposed work does not suffer from analog variations, nonlinearity caused by noise and does not require overhead of ADC/DAC. The regular and flexible structure of the compute array significantly improves energy and area efficiency when compared to previous digital accelerators.

Results and significance for the tinyML community

The proposed compute in-memory (CIM) bitcell architecture was fabricated using CMOS technology, tested and published (ESSCIRC '19). In addition to the published design, we have also designed Annealing Processor (to be presented at ISSCC '19) and another CIM bitcell array with analog voltage accumulation (to be presented at A-SSCC '19) based on the same circuit architecture.

We believe that our work can be a good fit for edge devices in energy constrained environments. The inherent robustness of the digital core provides an attractive solution to difficulties in deploying analog based accelerators. Also, since the importance of per-layer quantization was discussed in 2019 tinyML poster session ("Quantization for Efficient Inference in Edge Devices," R. Krishnamoor, Facebook), our work can be demonstrated as the hardware implementation (or extension) of initial discussion from last year.

50 Nanyang Ave, 639798, Singapore
+65-9456-0416 (Hyunjoon)

Emails:
kimh0003@e.ntu.edu.sg
bjkim@ntu.edu.sg