Modelling of Neural Network Hardware Accelerators

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Author: Ing. Jan Klhůfek Supervisor: Ing. Vojtěch Mrázek, Ph.D.

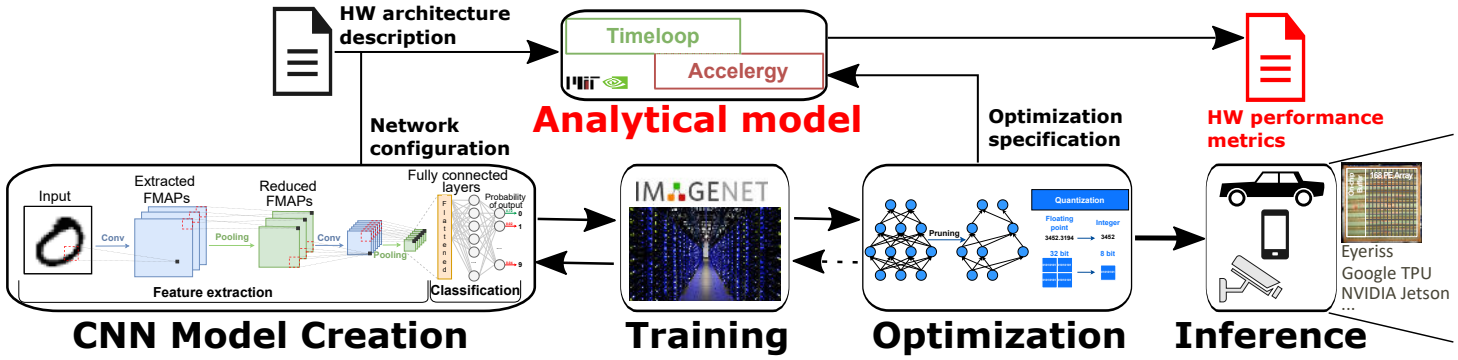
Motivation

Today's Convolutional Neural Networks (CNN) contain tens of **millions** to **billions** of **parameters**. This complexity necessitates huge number of computations – up to trillions of FLoating-point Operations Per Second (FLOPS). As we strive for real-time processing in embedded devices like smartphones, the demand for integrating these models is increasing.

Tackling the challenge of integration of a CNN model onto the target hardware (HW) is far from trivial. It requires meticulous model **optimization** to reduce both the number of computations and **memory size**, but most importantly, these optimizations must be done with the target hardware architecture in mind. The focus of this work is on the analytical modelling of mapping a workload onto a hardware accelerator, with a particular emphasis on supporting **quantization** as a way to optimize the memory utilization of the model onto the hardware. This opens up new possibilities for exploring the synergy between model optimization and HW deployment.

Aim of the thesis

The goal of this work was to extend a state-of-the-art (SOTA) analytical model to support **hardware quantization**, a technique that **reduces memory footprint** and **transfers**, leading to energy savings. Using such a model is orders of magnitude faster than real hardware inference, which enables feasible evaluation of HW metrics for various quantization strategies. Quantization maps infinite values to a finite set of values (floating-point to integer), optimizing memory use by using lower bitwidth and reduced precision – thus **increasing** the **hardware performance** and **reducing** the overall **energy consumption**. In order to achieve this, it is necessary to diligently map the CNN into the HW and schedule its operations in time. Energy savings are achieved by lowering the overall accesses to higher memory levels, such as the off-chip DRAM, which consumes over two orders of magnitude more energy per access.



Implementation results and future work

The extension's implementation utilizes a **bit-packing** technique, which effectivelly allows to store as many data elements into memory word as the word bits allow. Figure 1 illustrates the application of this technique in deploying the AlexNet CNN on a fixed HW architecture. By merely adjusting the precision of weights from 16 bits to 4 bits, we achieve **energy savings** of up to **27%**. Notably, the reduction in the hardware's energy consumption can be observed primarily in the memories responsible for weight storage. In future work, the exploration of more optimizations and the use of flexible hardware architectures that could influence computational units in addition to memory could be possible.

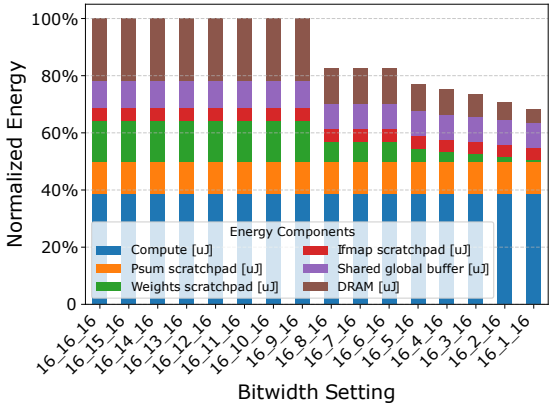
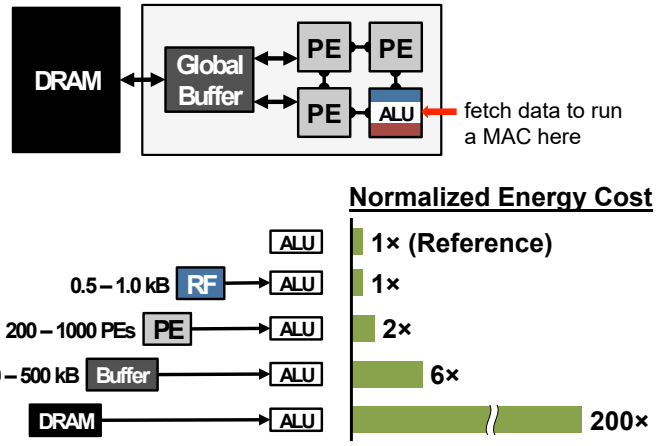


Figure 1: Impact of model's weights quantization on HW energy consumption



References and contributions

We extended the SOTA tool Timeloop [1] to enable mixed-precision quantization modelling in HW, intended for research purposes. The impact of this work on a real use case will be presented in [2].

[1] A. Parashar *et al.*, "Timeloop: A systematic approach to dnn accelerator evaluation," in *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2019, pp. 304–315.

[2] J. Klhufek, *et al.*, "Exploiting Quantization and Mapping Synergy in Hardware-Aware Deep Neural Network Accelerators " [In preparation]