Implementation of Fully-Pipelined Convolutional Neural Network Inference Accelerator on FPGA and HBM2 Platform

Name: NGUYEN VAN CAM
Laboratory: Computing Architecture
Supervisor: Professor Yasuhiko Nakashima

Abstract:

Due to the high speed and energy efficiency of the field-programmable gate array (FPGA), many FPGA-based inference accelerators for deep convolutional neural network (CNN) have been widely adopted. Large-scale CNNs require intensive computations as well as a large amount of storage space and memory access. However, low bandwidth off-chip memory is main challenge for data transmission between external memory and FPGA-based CNN inference accelerator.

In this research, we develop the following to improve throughput and energy efficiency for the FPGA-based CNN inference accelerator. First, we use a high bandwidth memory (HBM) to significantly expand the bandwidth of data transmission between the external memory and the accelerator, contrasting with conventional DDR memory. Second, a fully-pipelined manner, which consists of pipelined inter-layer computation and a pipelined computation engine, the output feature maps of each layer in the accelerator are computed row-by-row and immediately feed as input to the next layer, is implemented to speed up the inference time. Finally, a multi-core architecture with shared dual-buffers, which broadcasts the same kernel parameters among cores, is designed to reduce external memory access and increase the reusability of kernel parameters loaded into on-chip buffers.

We design the CNN inference accelerator on the Xilinx Alveo U280 platform with Verilog HDL and explore the VGG-16 model to verify the system during our experiment. With a similar accelerator architecture, the experimental results demonstrate that the memory bandwidth of HBM is 13.2× better than DDR4 memory. Compared with other accelerators in terms of throughput, our accelerator is 1.9×/1.3×/3.3× better than FPGA+HBM based/low batch size (4) DDR-based GPGPU/CPU. Compared with the previous FPGA+DDR based/DDR-based GPGPU/CPU accelerators in terms of energy efficiency, ours is 1.4-6.5×/1.6-4.2×/6.6× better, respectively.