Implementation of a Convolutional and a Pooling Layer of a CNN on FPGA

By Sina Mahdipour Saravani

sinamahdipour@aut.ac.ir

Abstract

FPGA-based neural network accelerators are recently becoming a valuable research topic. Although FPGAs have fewer resources than graphics processing units (GPUs), an efficient and intelligent FPGA implementation can provide the same throughput of a GPU with much less energy consumption. With FPGAs, the designer can create a specific hardware with an optimized architecture for the targeted neural network.

On the other hand, Convolutional Neural Networks (CNN) are among the most important and promising image processing techniques. The need and interest in applying CNN inference in embedded systems, like mobile devices, wearable gadgets, and automatic cars and aircrafts, has increased so much recently. In such systems, low energy consumption and low latency are the most important factors. Therefore, we want to accelerate the convolutional and pooling layers of a CNN using a FPGA chip.

In this bachelor thesis, we aim to efficiently implement the convolution and the max pooling functions of a CNN on a ZYBO FPGA board using the high-level synthesis tool by Xilinx. In our architecture, 3 filters with the size of 3*3 are applied to the input image matrix and then a max pooling function is applied to their outputs. This is accomplished by designing an IP core which puts the output matrices on its AXI ports.

We tested this IP core by both C/RTL co-simulation and on-board execution and the results showed the correctness of the IP's functionality. Also, we could successfully accelerate the convolution and max pooling functions relative to the software code on Central Processing Unit (CPU).

Keywords:

FPGA, Convolutional Neural Network, IP Core, Convolution, High-Level Synthesis, Pooling, ZYBO