FPGA based Convolutional Neural Network Acceleration

Yanchen Liu (yl4189), Minghui Zhao (mz2866)

February 26, 2022

1 Introduction

Convolutional Neural Network (CNN) is widely used in the machine learning task in the computer vision and neural language processing area. However, traditional general processing unit such as CPU can not provide ideal resource and allocate reasonably. GPU provides parallel computing by applying single instruction multiple threads on thousands of stream processor (core) and using Fast Fourier Transform (FFT) to accomplish fast convolution calculation. Although the acceleration of GPU is impressive, the rate is still limited by the instruction decoding, executing and memory sharing defined by the Von Neumann architecture. FPGA as a no instruction and no shared memory architecture device, is not limited by the rules and able to make pipeline and data paralleling at the same time. In this project, we plan to program the FPGA to accelerate a pre-trained CNN-based network for object detection and display the result on a monitor.

First, a real-time video stream will be captured by a Raspberry Pi + Camera, and sent via Ethernet to the HPS on DE1-SOC. The HPS which will then feed video stream to the CNN-based network via the HPS-FPGA high speed channel. After the FPGA finishes computation, the HPS obtains the output, then display the image frames with the bounding box on a monitor through VGA port.

2 Method

In this project, we plan to implement three popular approaches to accelerate a CNN object detection algorithm as follows.

2.1 FFT

The convolution of two matrices in the space domain equals to point-wise multiplication of two matrices in the frequency domain. The FPGA can achieve 2D fast Fourier transform and inverse fast Fourier transform efficiently with the parallel pipeline.

2.2 Img2col

The principle for Img2col algorithm is to transfer the convolution operator to the matrix multiplication, and do the matrix multiplication acceleration using the existing linear algebra algorithm (such as QR decomposition and Cholesky decomposition).

The feature matrix is flatten transferred to another matrix based on the kernel size and the kernel is flatted to a 1-D vector, then multiply the transferred matrix and the kernel vector to get a 1-D output. The final output is the reshaped 1-D output based on the feature and the kernel size. Description below is an example for the single channel Img2col, for multi-channel scenarios, the kernels are flatten to a matrix and all the other remain the same.

2.3 Winograd

Winograd is a popular convolution acceleration method for small kernel and title neural network. The basic idea of Winograd is similar with the FFT, to project the original data to another domain and do simple calculation then do the inverse transform. It improves the rate by reducing the times of multiplication and replacing them with the adding operation.

3 Milestone

1. Image data transmission between HPS and FPGA.
2. Acceleration algorithm FFT, img2col and Winograd.
3. VGA display with the bounding box.