ECGR 6090 Heterogeneous Computing
Baseline Implementation of MobileNetV1 on FPGA and GPU

Aneri Sheth	Saisha Kamat	Ushma Bharucha
801085402	801073710	801031321
asheth2@uncc.edu	skamat1@uncc.edu	ubharuch@uncc.edu

April 10, 2019

Summary

Increasing the accuracy of Convolutional Neural Networks (CNNs) has become a recent research focus in computer vision applications. Smaller CNN architectures like SqueezeNet and MobileNet can demonstrate accelerated performance on FPGAs and GPUs due to smaller model size and less network parameters. Implementation of CNNs on accelerators have two important benefits - GPUs provide thread-level parallelism to achieve higher throughput and FPGAs offer a customizable application-specific datapath. These two reasons make these platforms better suited for convolution like operations which involve huge data. This project aims to implement one such CNN architecture, MobileNet on an Image dataset in OpenCL, thereby comparing kernel execution time and memory bandwidth usage on FPGA and GPU.

Project Details

- CNN architecture will be performed on 1 image (from any image dataset) and existing training model weights will be taken.
- Type of Optimizations - source-level optimizations - Xilinx Pragma, OpenCL pipes
- Implementation Plan - Implementing MobileNet CNN architecture in OpenCL which consists of 28 layers involving Maxpooling, ReLU, standard convolution and depthwise separable convolution (depthwise and pointwise).
- Targeted Platforms - Nvidia K80 GPU (Mamba cluster) and AWS Xilinx FPGA

Achievable Outcomes

- Baseline implementation of MobileNet architecture in OpenCL for FPGA and GPU
- Comparing performance of baseline implementation on FPGA and GPU based on kernel execution time and memory bandwidth usage
- Trying out source level optimizations like Xilinx Pragma and OpenCL pipes and compare it with baseline.

Individual Responsibility

Aneri	Depthwise separable convolution, ReLU, run on FPGA, OpenCL pipes
Saisha	Standard convolution, Maxpooling, run on FPGA, OpenCL pipes, Xilinx Pragma
Ushma	Run on GPU, Kernel execution time, memory bandwidh comparison on FPGA and GPU, OpenCL pipes