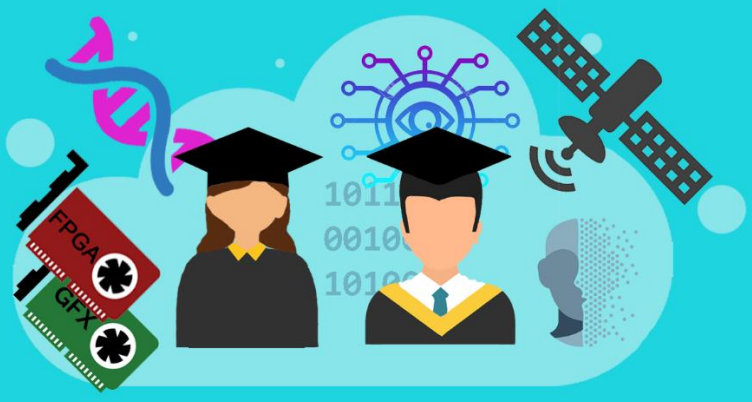
Diploma Thesis

Microproccessors and Digital Systems Laboratory

Academic year 2020-2021



Building optimized FPGA accelerators for CNN prediction and comparison with GPU accelerators

Machine Learning is a thriving area in the field of Artificial Intelligence, which involves algorithms that can learn and predict through building models from given datasets for training. Many machine learning designs are based on deep learning networks, which involve an architecture with neurons grouped into layers by their functionalities, and multiple layers organised to form a deep sequential structure. Convolutional Neural Network (CNN) is a classic deep learning network which has been applied to many vision-based tasks. There are four well-known CNN architectures in recent years, AlexNet, VGGNet, Inception, ResNet.

The accuracy of convolutional neural networks (CNNs) has been continuously improving but the computational cost of these networks also increases significantly. Real-world systems may suffer from the low speed of these networks. For example, a cloud service needs to process thousands of new requests per seconds. It is thus of practical importance to accelerate test-time performance of CNNs. Many hardware technologies can be used in accelerating machine learning algorithms, such as Graphics Processing Unit (GPU) and Field-Programmable Gate Array (FPGA). FPGA still remains less popular than GPU and CPU regarding CNN model deployment platform, mainly due to the difficulty that lies in converting high-level CNN descriptions to runnable FPGA hardware designs.

The scope of this diploma thesis is to build efficient CNN accelerators for FPGAs, incorporate them into widely-used toolflows and compare them against GPU implementations. An open-source OpenCL-based FPGA accelerator will be adopted to build optimal accelerators for both popular ((AlexNet, VGGNet, ResNet) and user-defined networks. The accelerator will be built based on a tuning process and optimization strategy, that takes into account the target FPGA platform. The generated accelerators will instantiate pre-trained models available in Tensorflow and Keras libraries and will therefore be integrated in workflows (e.g. written in python) utilizing these models. These workflows will also be evaluated through the use of a Tensorflow release with GPU support. Therefore, the thesis will demonstrate the translation of various model architectures to FPGA designs and also produce a comparative study for inference performed by built-in tensorflow prediction methods (CPUs), FPGAs and GPUs. The resulted workflows can be further tested on a cloud-testbed with access to GPU and FPGA devices.

PREREQUISITES: OpenCL programming, C/C++ Programming, Python, Scripting Skills

Academic Advisor1 : Konstantina Koliogeorgi, konstantina@microlab.ntua.gr
Academic Advisor2 : Sotirios Xydis, sxydis@microlab.ntua.gr
Academic Advisor3: Dimitrios Soudris, dsoudris@microlab.ntua.gr