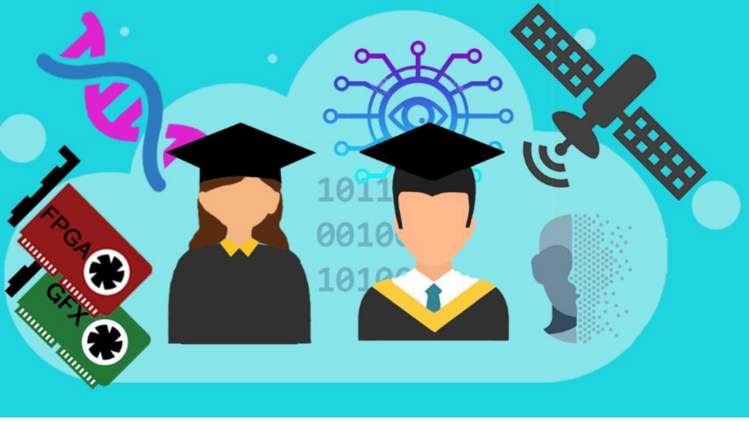
Diploma Thesis

Microproccessors and
Digital Systems
Laboratory



Design and Performance-Power Analysis of Multi-Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs) on FPGAs through Multiple Design Flows

The increasing demands for high-performance Machine Learning (ML) models has led to the utilization of hardware accelerators (GPUs, FPGAs) for the execution of their inference phase.

The design and acceleration of complex ML networks on FPGAs is not always a straightforward task. Different design architectures that trade-off performance to power consumption can be evaluated before the final design, whereas the typical Register-Transfer Level (RTL) design flow cannot be used in complex designs due to the time that is required for the development.

Even High-Level Synthesis (HLS) methodologies that are used for the hardware design of customized architectures are not always feasible in complex ML models. For that reason, python-based frameworks that expand the capabilities of known Python ML frameworks (e.g pyTorch, TensorFlow) to a hardware level have been introduced. The most common open-source frameworks are FINN and hls4ml.

In the context of this thesis different ML models (MLPs and CNNs) will be developed following three design flows: **1)** a custom HLS architecture, **2)** a dataflow architecture using FINN, **3)** a fully parallel architecture using hls4ml. The different MLP and CNN designs will be evaluated for their performance, power efficiency and the design time on high and low-end FPGAs

RELATED MATERIAL:

- Umuroglu, Yaman, et al. "Finn: A framework for fast, scalable binarized neural network inference." *Proceedings of the 2017 ACM/SIGDA international symposium on field-programmable gate arrays*. 2017
- Fahim, Farah, et al. "hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices." *arXiv preprint arXiv:2103.05579* (2021).
- Ngadiuba, Jennifer, et al. "Compressing deep neural networks on FPGAs to binary and ternary precision with hls4ml." *Machine Learning: Science and Technology* 2.1 (2020): 015001.

- https://xilinx.github.io/finn/
- https://github.com/fastmachinelearning/hls4ml-tutorial

CONTACT INFORMATION:

- Argyris Kokkinis, Ph.D student.: (arkokkin@auth.gr)
- Associate Prof. Kostas Siozios: (ksiop@auth.gr)