

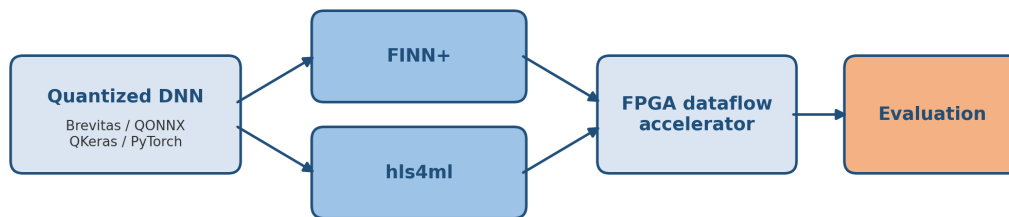
Comparing DNN Dataflow Compilers: FINN+ vs. hls4ml

Bachelor Thesis

At a glance

- Get familiar with the FINN+ and hls4ml tool flows
- Define a benchmark suite of quantized DNNs (e.g. MLP, CNN, Transformer)
- Build and deploy FPGA accelerators with both compilers
- Systematically evaluate and compare both compilers

Deep neural networks (DNNs) are increasingly used in applications with demanding performance and efficiency requirements, from trigger systems in particle physics to datacenter inference. FPGAs are well suited for such workloads, in particular when networks are quantized to low bit widths and mapped onto a custom dataflow architecture, in which every layer is implemented as its own pipelined hardware stage. To make this technology accessible without in-depth hardware design expertise, dataflow compilers translate trained, quantized networks directly into FPGA accelerators.



Two prominent open-source frameworks follow this approach. FINN+, an extended fork of AMD's FINN framework maintained at Paderborn University, compiles quantized networks from Brevitas/QONNX into streaming dataflow accelerators and adds, among others, support for Transformers, Alveo datacenter cards and multi-FPGA deployment. hls4ml, originating from high-energy physics, translates Keras/QKeras, PyTorch and ONNX models into HLS code with a focus on ultra-low latency. Although both tools address similar use cases, they follow different design approaches and generate quite different hardware architectures. A systematic side-by-side comparison of the two is still missing.

In this thesis, you will set up both tool flows and select representative benchmark networks, such as MLPs, small CNNs or Transformers. You will compile each network with both frameworks under comparable configurations, deploy the resulting accelerators on AMD FPGAs and evaluate them systematically. Beyond the generated hardware, the comparison should also consider the capabilities and practical use of both frameworks. The results will reveal the strengths and limitations of each compiler and show which tool suits which scenario.

Further reading:

<https://github.com/eki-project/finn-plus>

<https://github.com/fastmachinelearning/hls4ml>

Contact:

Linus Jungemann, Phone: +49 5251 60-4635 E-Mail: linus.jungemann@uni-paderborn.de