

FINN+: Highly Efficient Neural Network Inference in the Datacenter

Authors:

Bjarne Wintermann, Linus Jungemann

Co-authors:

Felix Jentzsch, Christoph Berganski

Due to increasing AI usage during the recent years, acceleration of both training and inference for deep neural networks has become an important topic. While training represents a large up-front effort, permanently running the model for large groups of users requires significant compute as well. However not only does inference need to be fast, both in latency and throughput, but also energy efficient.

Neural networks lend themselves to being accelerated by dedicated hardware, since many architectures share the same basic operations. One specific accelerator type is the Field Programmable Gate Array (FPGA). FPGAs can enable optimized model-specific designs, and consequently have good performance and low energy requirements. Additionally, they are reconfigurable and thus avoid being locked into a specific architecture, which is a drawback of ASIC inference accelerators.

We present FINN+, a fork of FINN which is being co-developed by the Computer-Engineering Group (CEG) and the Paderborn Center for Parallel Computing (PC2) at Paderborn University. FINN itself is one of the largest frameworks for neural network inference on FPGAs. The research project was originally developed and published by AMD / Xilinx in 2017, and has since become a well known open-source framework still under active development by both AMD and a large community of researchers.

FINN+ aims to further extend FINN in multiple directions. One such extension is the high-performance C++ driver. Utilizing host-memory access, improved data-packing algorithms and various compile-time optimizations, the driver facilitates the usage of FINN-built designs in the datacenter. This, for example, allows it to compete with an A100 GPU on certain models and batch sizes [1].

To accommodate large models, FINN is able to „fold“ designs: This lowers resource usage, but also decreases performance. To enable increasingly large models in FINN and avoid unnecessary folding, we added a Multi-FPGA extension which allows FINN+ to stream data between parts of the model distributed to interconnected FPGAs [2]. Furthermore, FINN+ supports Transformers on FPGAs. This functionality was demonstrated on small GPTs, showcasing text generation abilities [3]. In addition, several more features, such as an improved streamlining flow and empirical QoR estimation are actively being developed.