DEPLOYING & ACCELERATING DL MODELS ON FPGAs

DESIGNED BY

- Muhammad Abdullah Khan
- Muhammmad Junaid Ali
- Amur Saqib Pal
- Sannan Zia Abbasi

NUST
SCHOOL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCE

Introduction

Our goal is to accelerate artificially-intelligent models on specialized hardware for improved real-time inference.

Results

1000% faster inference than CPU, minimal loss in accuracy.

We obtained the following results. Note the accuracy drop on the FPGA as compared to the GPU.

- MNIST (10 classes, 96.77% accuracy on GPU, 96.24% accuracy on FPGA BNN)
- EMNIST-Digits (10 classes, 97.86% GPU, 97.62% BNN)
- EMNIST-Letters (26 classes, 84.85% GPU, 80.7% BNN)
- EMNIST-Letters+Digits (62 classes, 76.86% GPU, 70.9% BNN)

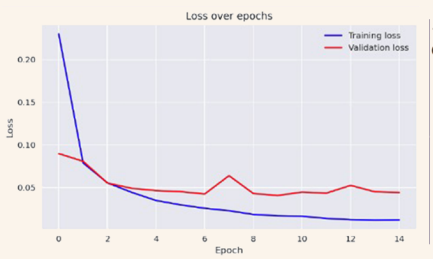
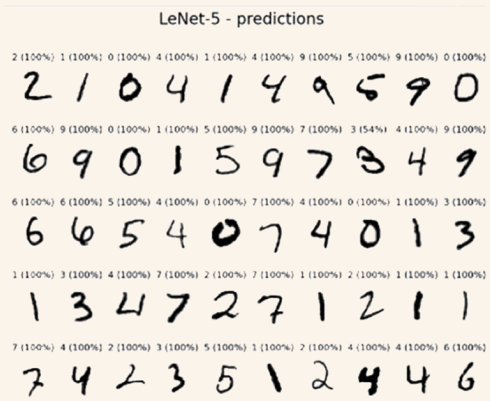


Table 2. Energy consumption of multiply-accumulations (Horowitz, 2014)

Operation	MUL	ADD
8bit Integer	0.2pJ	0.03pJ
32bit Integer	3.1pJ	0.1pJ
16bit Floating Point	1.1pJ	0.4pJ
32tbit Floating Point	3.7pJ	0.9pJ

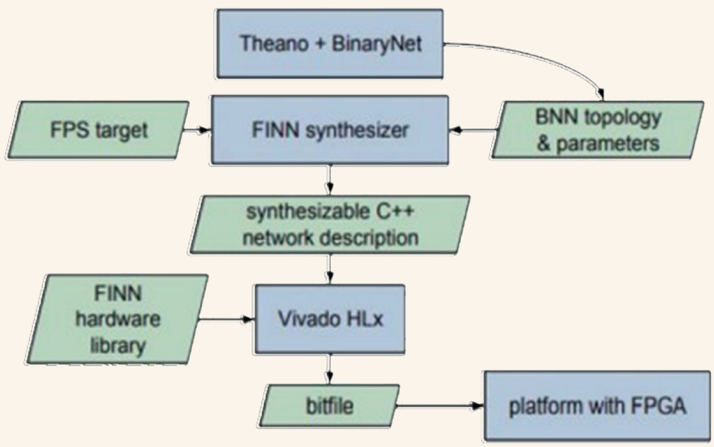
Table 3. Energy consumption of memory accesses (Horowitz, 2014)

Memory size	64-bit memory access
8K	10pJ
32K	20pJ
1M	100pJ
DRAM	1.3-2.6nJ



Methodology

We used optimization techniques to break down complicated PyTorch DNNs into simpler and more efficient models so to bear fruitful results when we run them on an FPGA, namely the PYNQ Z1. These optimization techniques include quantizing and binarizing the weights, making "tiny" equivalents of the complex models, and more. Moreover, high-level synthesis tools such as Vitis IDE and Vivado HLS are used for hardware implementation.



Conclusion

Our method of implementing NNs can be used for numerous real-life applications where classification, segmentation, and other AI-related tasks are to be performed

We trained a simple Binarized MLP (1 FC with 64 Neurons) with 93% accuracy on MNIST and ran it on 4 different platforms with maximum PE and SIMD count and achieved the results we were expecting.

Device	Clock (MHz)	FPS	Accuracy (%)
ZYNQ ARM Processor	650	710	89
ZYNQ FPGA	100	**6.1M**	89
Intel CPU	1900	60k	93
NVIDIA Tesla K20	758	600k	93

Software Used

References

1. FINN: A Framework for Fast, Scalable Binarized Neural Network Inference
2. Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or -1