Neural Network Overlay Using FPGA DSP Blocks

Lenos Ioannou and Suhaib A. Fahmy

School of Engineering, University of Warwick, UK

WARWICK
THE UNIVERSITY OF WARWICK

Introduction

WARWICK
THE UNIVERSITY OF WARWICK

- Long back-end tool compilation hinders rapid deployment of Neural Networks on FPGAs at the edge
- Use of overlays to build abstractions on top of the FPGA
  - Effectively enabling rapid deployment
- Core NN operation, multiply-accumulate, maps well to DSP Blocks
- Most FPGA NN implementations operate sub-max frequencies [1]
  - Can be solved by optimising the overlay around the DSP blocks [3]

Neural Network Test Cases

WARWICK
THE UNIVERSITY OF WARWICK

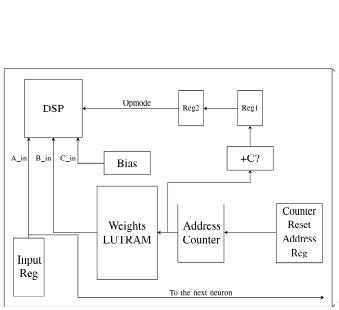
- Trained 3 NNs using Tensorflow [2], each one comprises four layers
- Use of ReLU in the intermediate layers

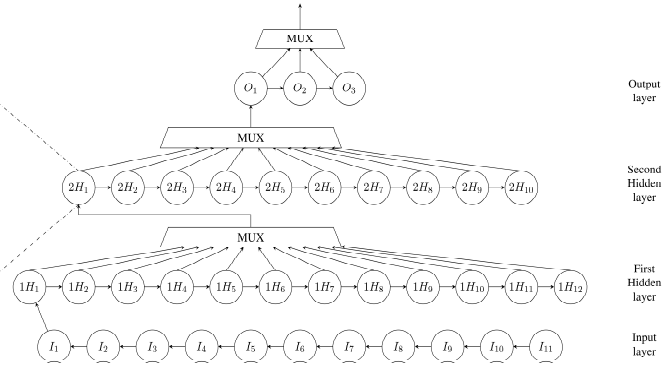
Dataset	NN Topology	Acc. Train	Acc. Test
Customer Churn Dataset	11-6-6-1	84.26%	82.95%
Diabetes Dataset	8-12-8-1	78.39%	-
Iris Dataset	4-10-10-3	98.33%	96.67%
Overlay	11-12-10-3	-	-

- Considering the input bit-widths of the DSP48E2:
  - 18 bit weights
  - 27 bit inputs
  - 48 bit biases

Overlay

- Each neuron is mapped to a single DSP block
- DSP blocks alternate between two opmodes
- Serial data flow
  - Needs to stall when # neurons > # inputs
- Adjustable latency





Implementation Results

WARWICK
THE UNIVERSITY OF WARWICK

- Implemented the overlay targeting the **Zynq Ultrascale+ ZU7EV**

LUTs	LUTRAM	FFs	DSPs	Frequency (MHz)
796	225	2552	25	770

- Maintains low resource utilization
  - Feedforward serial data flow is highly efficient
- High operating frequency
  - Near the DSP blocks' theoretical maximum

WARWICK
THE UNIVERSITY OF WARWICK

Conclusion

- Not offering peak performance in a particular NN implementation
- Contribute to the more rapid deployment of NNs on FPGAs at the edge
- Prioritise low resource utilization and energy efficiency

Future work

- Implement a mechanism to handle the data flow and stall accordingly
- Expand the overlay for deeper topologies
- Integration with a rapid compiler flow

WARWICK
THE UNIVERSITY OF WARWICK

References

[1] E. Wu, X. Zhang, D. Berman, and I. Cho, “A high-throughput reconfigurable processing array for neural networks,” in Int. Conference on Field Programmable Logic and Applications (FPL), Sep. 2017.

[2] Martin Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[3] A. K. Jain, D. L. Maskell, and S. A. Fahmy, “Throughput oriented FPGA overlays using DSP blocks,” in 2016 Design, Automation Test in Europe Conference Exhibition (DATE), March 2016, pp. 1628–1633.

[4] A. K. Jain, X. Li, P. Singhai, D. L. Maskell, and S. A. Fahmy, “DeCO: A DSP block based FPGA accelerator overlay with low overhead interconnect,” in Proc. Int. Symposium on Field-Programmable Custom Computing Machines (FCCM), 2016, pp. 1–8.