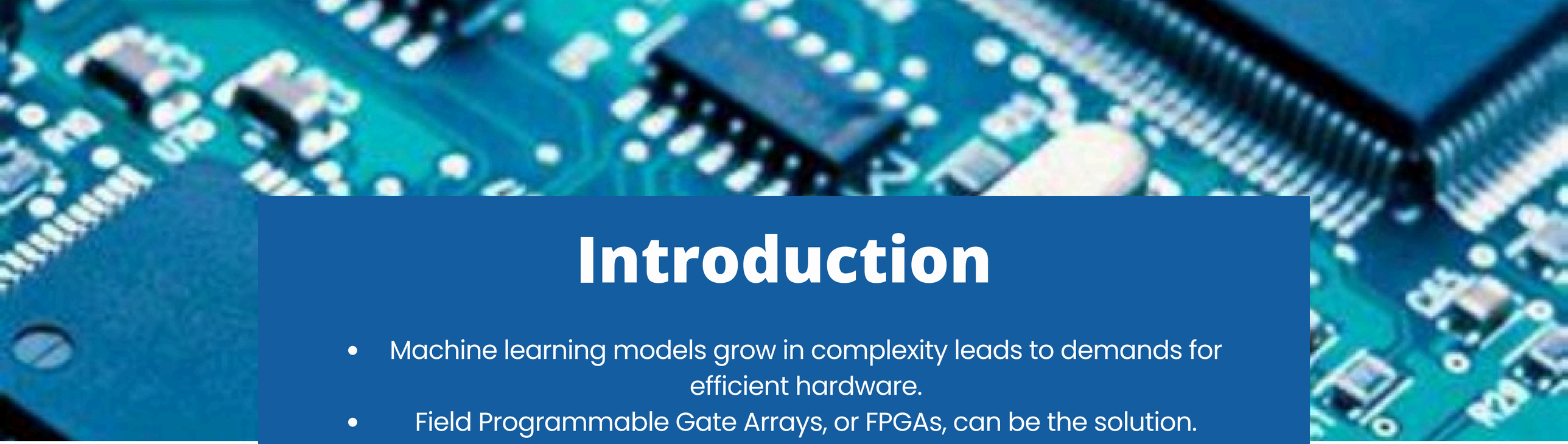
USYD

Effective Hardware Block Design Modifications: Impact on Neural Network Performance

Student: Prosper Su
Supervisior: Dr. David Boland



Introduction

- Machine learning models grow in complexity leads to demands for efficient hardware.
- Field Programmable Gate Arrays, or FPGAs, can be the solution.
- The goal is to investigate whether changes in hardware can improve models.

Research Question

How can hardware-level modifications be utilized as a quick and cost-effective approach to improve the performance of FPGA-based neural networks?

01

Understand the quantized neural networks by testing their accuracies and resource utilizations.

02

Design and implement direct modification techniques at hardware-level on FPGA-based neural networks.

03

Evaluate the effectiveness of FPGA tuning as a cost-effective and quick approach by conducting the performance comparisons.

Background

Software-based optimizations can be limited by the fixed configurations on hardware platforms.

Hardware modifications offer better adabbility and efficiency.

Brevitas

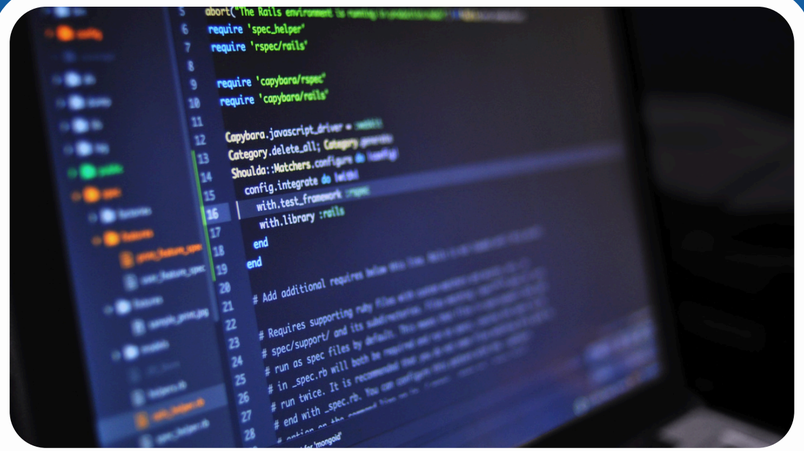
- PyTorch-based library for quantization-aware training.
- design for efficient and low-precision neural networks.

FINN

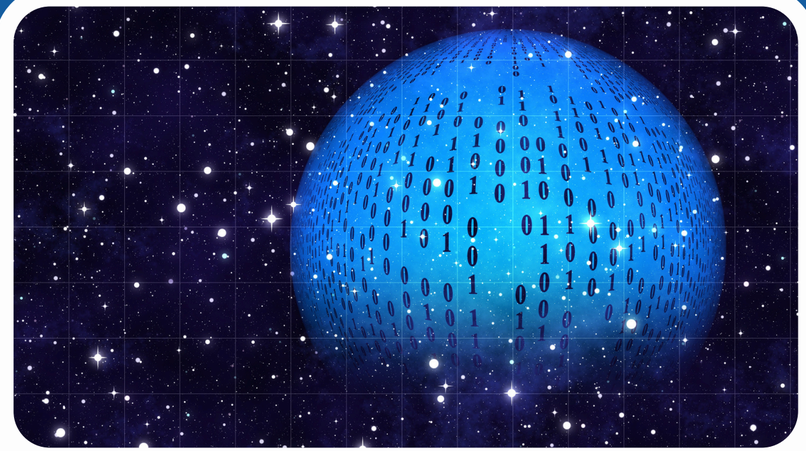
- A FPGA-optimized framework for hardware synthesis by Xilinx.

Methodology

Sample metrics testing

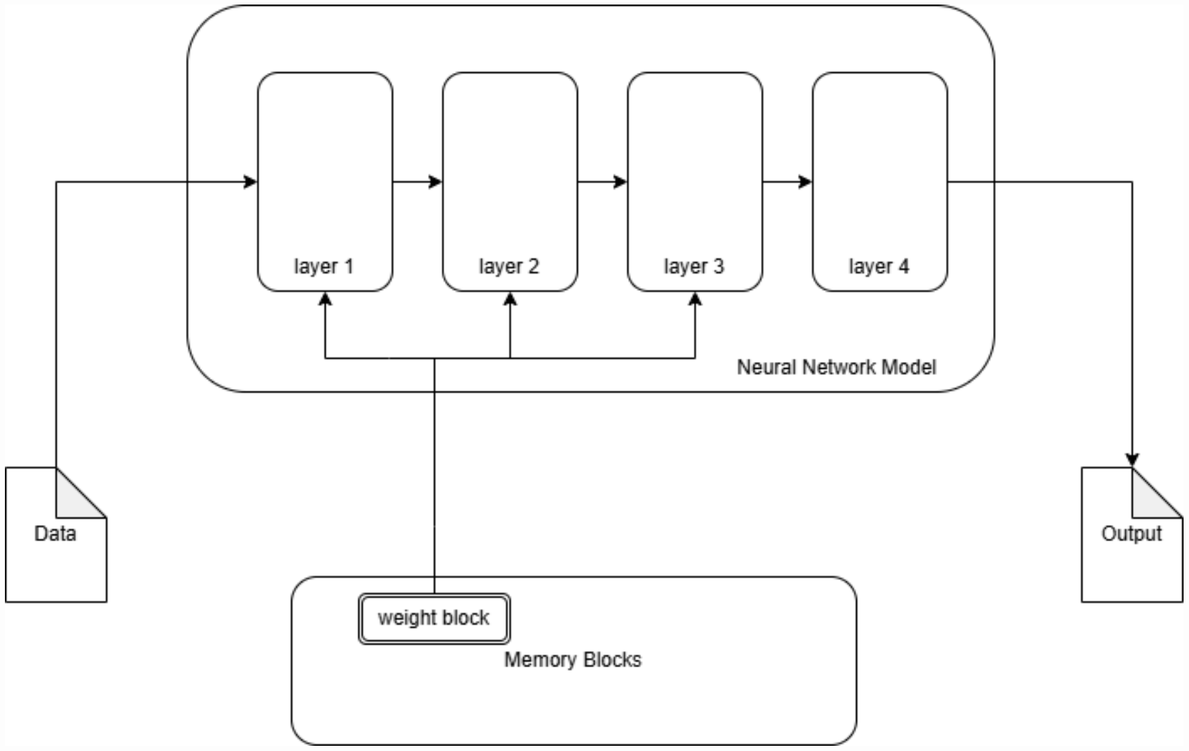


Accuracy testing with modification

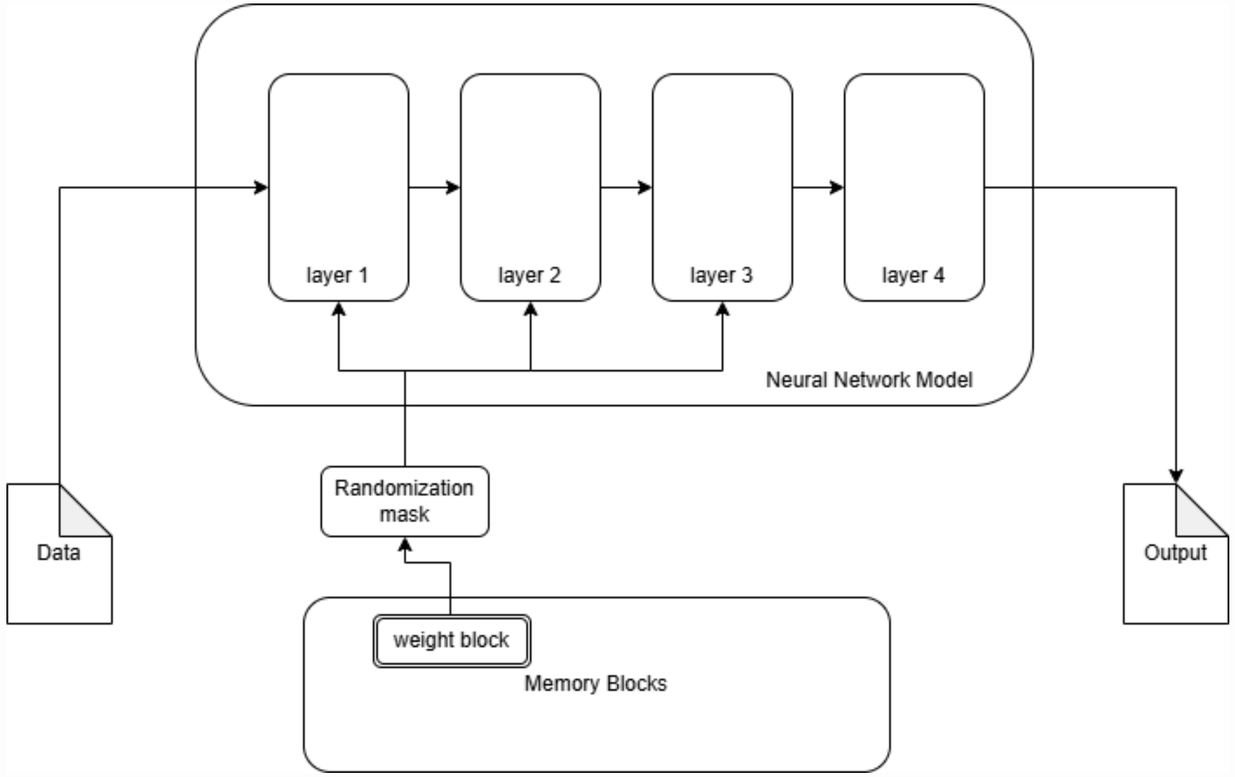


Resource utilization testing with modification

Accuracy testing

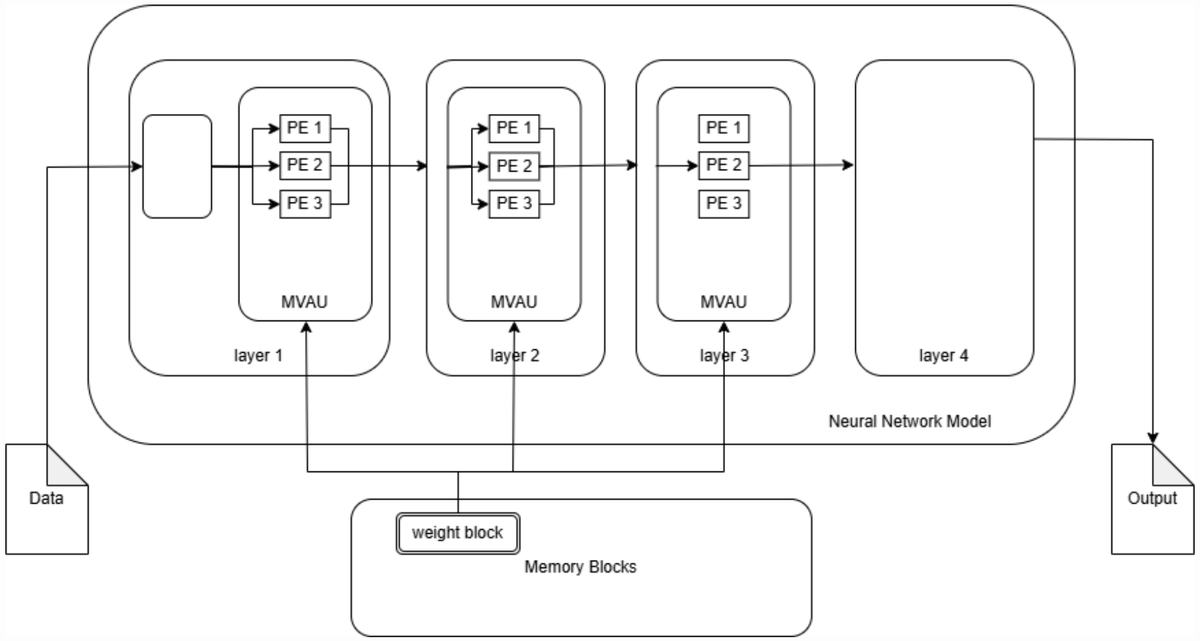


Default model

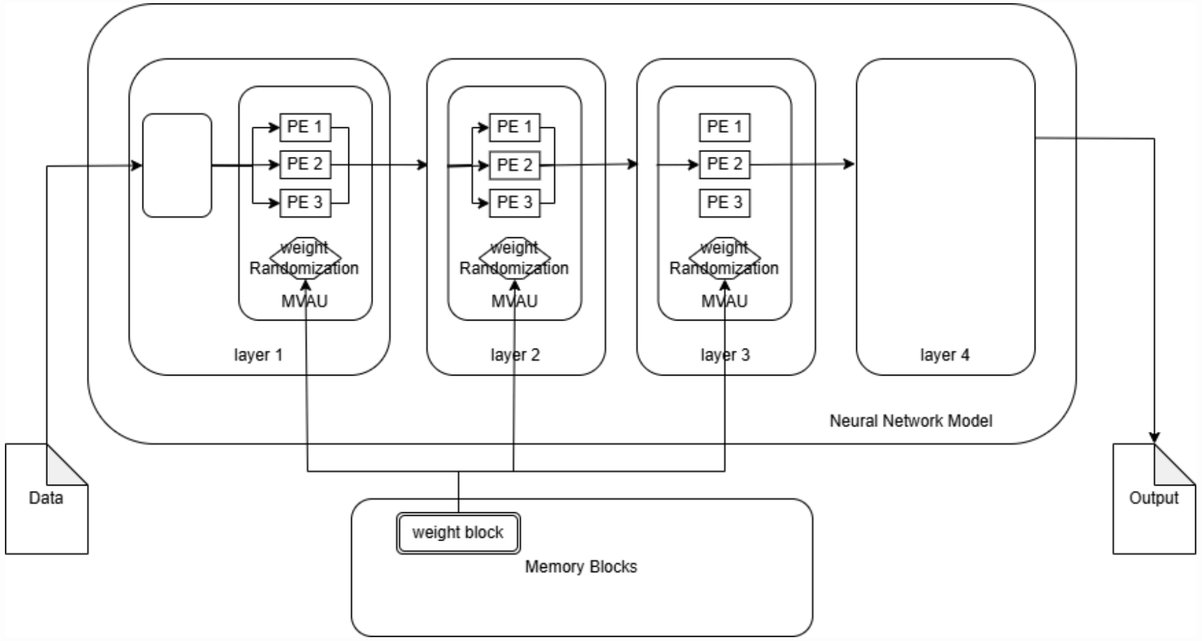


Model with modification in Brevitas

Resource utilization testing



Default model



Model with modification in FINN

RESULTS

The results are collected by the methods mentioned in the previous section.

Reference

with a default qunatized neural network

Accuracy result in Brevitas:

73%

Resource results in FINN:

6586	7711	22
LookUp Tables counts	Flip-Flops counts	Block ram counts

Testing results

after modifications applied

Accuracy result in Brevitas:

81.5%

Resource results in FINN:

7054	8043	22
LookUp Tables counts	Flip-Flops counts	Block ram counts

Comparisons

Accuracy increases by:

12%

Resource results in FINN:

7.1%	4.3%
LookUp Tables counts	Flip-Flops counts

```
test(model, test_quantized_loader)
```

```
0.730505757178254
```


```
{
  "vivado_proj_folder": "/tmp/finn_dev_su/synth_out_of_context_xjeubqhy/results_finn_design_wrapper",
  "LUT": 6586.0,
  "FF": 7711.0,
  "DSP": 0.0,
  "BRAM": 22.0,
  "BRAM_18K": 0.0,
  "BRAM_36K": 22.0,
  "URAM": 0.0,
  "vivado_version": 2022.2,
  "vivado_build_no": 3671981.0,
}
```

Discussion

From this experiment, several insights are found:

Trade-offs

- The increase in accuracy leads to an increase in resources.

Significance of Hardware Modifications

- Direct hardware modifications can enhance performance without significant additional cost.

Conclusion

- Hardware-level modifications can be a cost-effective, quick way to improve performance.
- Accuracy improvements without substantially increasing resource usage.

Future work

- dynamic quantization that adjusts based on model demands to further control resource use.
- deploying these modifications on physical FPGA hardware would validate their impact on performance factors.

Q & A

THANK YOU!