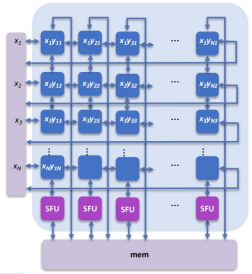
Weight- and output-stationary reconfigurable 2D systolic array-based AI accelerator and mapping on Cyclone IV GX — ECE 284 Group10

Group Member: Zhongdongming Dai, Jheng-Ying Lin, Kejia Ruan, Yanghe Sun, Hongjie Wang

- Summary

In order to improve the efficiency and versatility of the 2D systolic array hardware, our group present a reconfigurable approach, with several additional alphas.



- Experimental Results

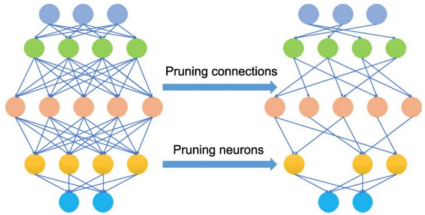
Quantization Aware Training

	VGG16	Resnet20(alpha)
Accuracy	92%	88%
Quant Error	2.257e-07	0.7054

Mapping on FPGA (Cyclone IV GX)

Cyclone IV GX FPGA Mapping of Vanilla Version	
FMax(MHz)	127.99
OP	128
GOPs/s	16.38
TOPs/W	3.76e-09
Core Dynamic Power(mW)	34.01
Logic Elements	22,414
Registers	12,098

- Alpha1 Unstructured and Structured Pruning VGG16



	Sparsity	Acc before fine-tuning	Acc after fine-tuning
Unstructured	0.8	10%	89%
Structured	0.8	10%	79%

- Alpha2 Quantization Aware Trained Resnet20 Mapping

```
layer = list(model.children())[4][1].conv1 # 2nd BasicBlock's 1st Conv2d Layer
layer
```

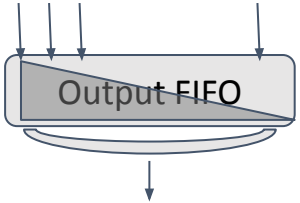
```
QuantConv2d(
  8, 8, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False
  (weight_quant): weight_quantize_fn()
)
```

Layer Info:
Input nij: 18*18
Output nij: 16*16
Kij: 3*3

Verification Process:
- P_Mem Address.txt needs 256(o_nij) * 9(kij) = 2304 lines and each line needs 12 bits to represent

- Alpha3 Concurrent OFIFO Read/Write

- Concurrently read/write OFIFO can reduce FIFO depth.
- Reduce required depth from 64 to 18, decreasing hardware complexity and costs.



- Acknowledgements

This project builds upon contents from Prof. Mingu Kang 's Course ECE 284, VVIP Lab UCSD 24 Fall