NANYANG TECHNOLOGICAL UNIVERSITY
School of Computer Science and Engineering

IEEE MMSP 2019

Lowering Dynamic Power of a Stream-based CNN Hardware Accelerator

Duvindu Piyasena, Rukshan Wickramasinghe, Debdeep Paul, Siew-Kei Lam and Meiqing Wu

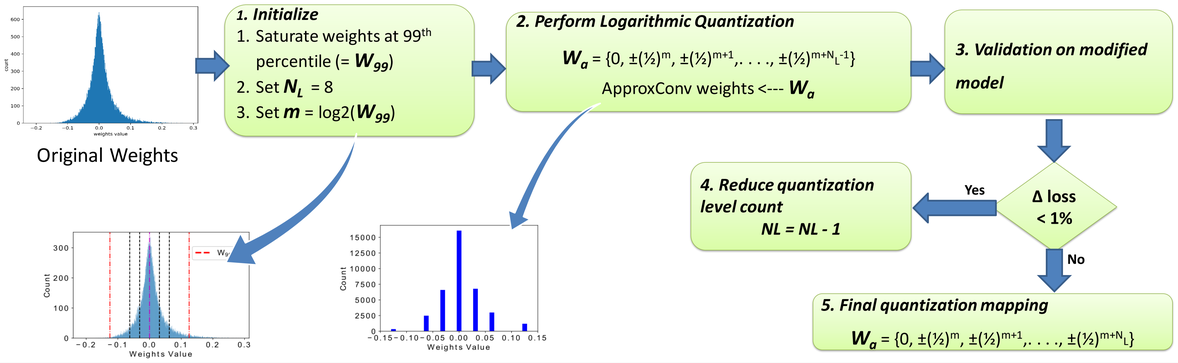
1. Motivation

- We exploit the redundancies occurring as a result of ***Max Pool (MAX)*** downsampling effect in CNNs and propose a method to eliminate the redundancies to save dynamic power in FPGA stream-based CNN accelerators
- $\%\ FLOPS\ redundancy\ in\ a\ CONV-MAX\ layer = \left(1 - \frac{1}{(Stride\ of\ MAX)^2}\right) * 100$

  eg : if Max pool stride = 2, FLOPS redundancy = 75%

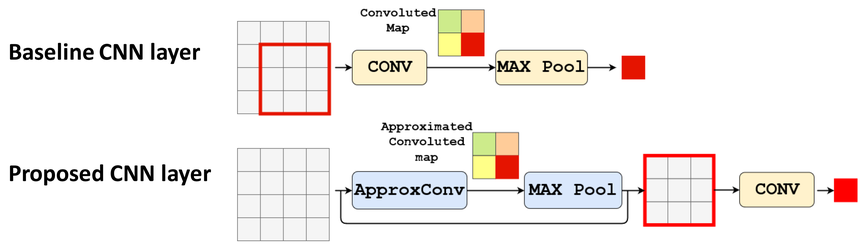
3. Approximation Scheme

The proposed ***ApproxConv*** performs Convolution operation with original CONV weights quantized to power-of-2 levels, which enables use of light-weight bit-shifters in place of costly multipliers. This is further optimized by performing a static analysis to identify the least number of quantization levels required (N_L) using an iterative search.



2. Proposed Method

The proposed method aims to eliminate the computational redundancies arising from Max pool layer by predicting the feature map candidates in the neighbourhood that will result in maximum activation prior to performing Convolution. This scheme is referred to as '***ApproxConv***'.



4. Accuracy Evaluation

a) Quantization level search

b) Accuracy Comparisons Compared with ***Signconnect*** proposed in previous work(*), which uses the sign of the weights to perform the approximations

TABLE II: Accuracy Evaluation

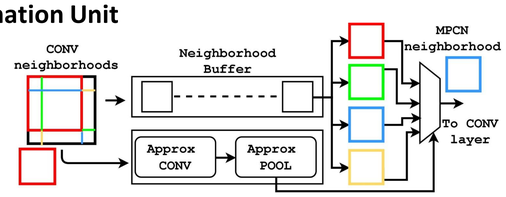
Network	Baseline Accuracy (Top-1/5)	Sign Connect	Proposed Method NI (By Layer)	Proposed Method Power-of-2 Levels (By Layer)	Proposed Method Accuracy (Top-1/5)
VGG16	68.15/88.14	67.99/87.78	3 2 2 2	0.0625, 0.125, 0.250 0.031250, 0.0625 0.015625, 0.03125 0.015625, 0.03125	68.02/88.09
AlexNet-BN	56.57/ 79.92	21.13/40.60	4 2 3	0.03125, 0.0625, 0.125, 0.25 0.03125, 0.0625 0.015625, 0.03125, 0.0625	56.11/79.37
Cifar10-Quick	72.19/97.70	70.88/97.81	1	0.125	71.87/97.69
Cifar10-Full	81.66/99.12	74.53/98.53	2	0.125, 0.25	81.42/99.07
Cifar10-NiN	89.57/99.62	89.43/99.62	2	0.25, 0.5	89.49/99.62
Lenet	99.01/99.99	99.05/99.99	2 1	0.5, 0.25 0.125	99.05/100

* T. Ujiie, M. Hiromoto, and T. Sato, "Approximated prediction strategy for reducing power consumption of convolutional neural network processor," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2016, pp. 870–876

5. Hardware Evaluation

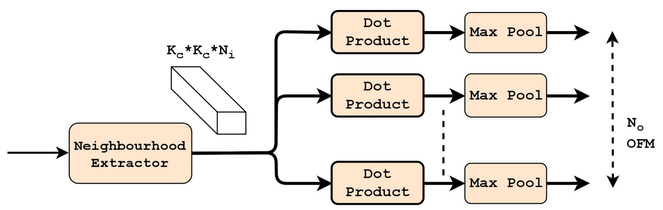
- Implementation done in VHDL based on *Haddoc2. (*)*
  - **Operating Frequency** : 100Mhz
  - **Device** : Xilinx Virtex Ultrascale+ xcvu9p
  - **Synthesize tool** : Xilinx Vivado 2018.3
  - **Simulator** : Mentor Modelsim 10.3
  - **Power Estimation Mode** : Post-Synthesis Timing Simulations
- Power Gains achieved by clock gating CONV circuitry via ApproxConv predictions

Approximation Unit

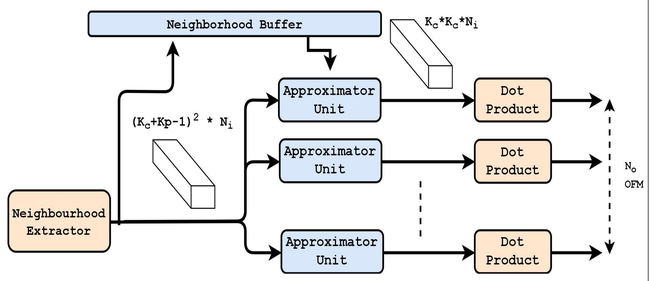


* K. Abdelouahab, M. Pelcat, J. Srot, C. Bourrasset, and F. Berry, "Tactics to directly map cnn graphs on embedded fpgas," *IEEE Embedded Systems Letters*, vol. 9, no. 4, pp. 113–116, Dec 2017

Baseline hardware architecture(single layer)



Proposed hardware architecture (single layer)

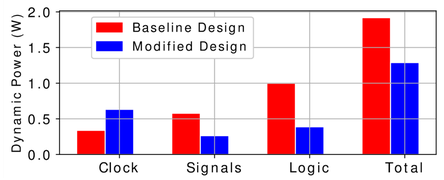


Hardware Evaluation Results

TABLE III: Hardware Evaluation Results

		Baseline	Modified	Change (%)
Dynamic Power (W)		1.919	1.289	**-32.83%**
Resource	LUT	431752	814558	88.66%
	FF	156178	317096	103.03%
Latency (ns)		7980	8010	0.38%
Energy/Image (J)		1.53E-05	1.03E-05	**-32.58%**

Dynamic Power Estimation Breakdown



6. Results Summary

- Power/ Energy gains : **33%**
- Latency change : **0.38%**