UC San Diego
Electrical and Computer Engineering
JACOBS SCHOOL OF ENGINEERING

Software-Driven Optimization of 4-bit Quantized VGG for Systolic FPGA Accelerators

ECE 284
Fall 2025

Elephant

Sujen Kancherla, Amaan Mohammed, Matteo Persiani, Rishi Pothukuchi, Daniel Sanei

Motivation

Our goal is to design, verify, and optimize a custom AI accelerator for CIFAR-10 using quantized VGG and systolic array hardware. We design an accelerator that is bit adaptive (2/4 bit), stationarity reconfigurable (WS / OS), and scalable 2D systolic using Cyclone IV FPGA.

Vanilla Model

Our vanilla model follows basic VGG16 architecture with 4-bit activations and weights. Per the specifications, we squeezed quantized layer 27 (channels to 8×8), and removed BatchNorm at the squeezed conv.

VGG16 Quantization Aware Training

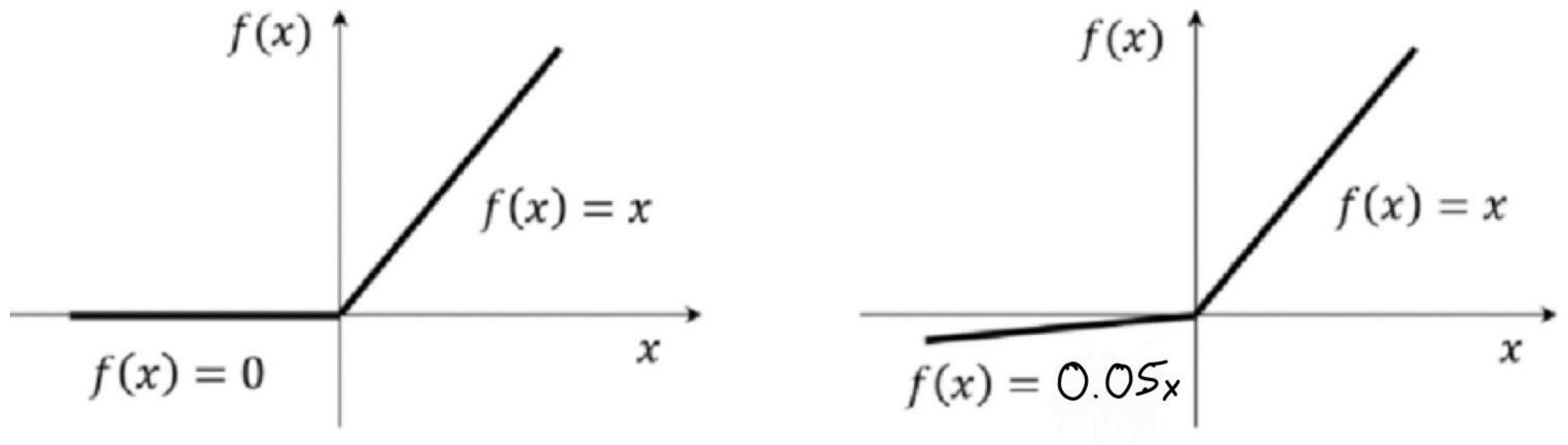
	VGG16 (Vanilla)	Alpha 1	Alpha 2	Alpha 4
Accuracy (CIFAR10)	90.33%	90.74%	91.15%	90.82%
Quantization Error	0.0000003776	0.0000004904	0.0000005847	0.0000007252

Mapping on FPGA (Cyclone IV GX)

	VGG16 (Vanilla)
Operations	128
Frequency	100.00 MHz Fmax at -40°C: 118.71 MHz
Dynamic Power	209.62 mW (0.20962 W)
Peak Throughput (GOPS/s)	12.8
Energy Efficiency (GOPS/W)	61.0
Logic Elements	17,348

Alpha 1 (Leaky ReLU)

- **Replacing ReLU with Leaky ReLU in our quantized VGG16 improved classification accuracy by +0.41% on CIFAR-10. This gain is probably due to improved gradient flow and reduced saturation in 8-channel layers.**
  - ○ **Hardware Impact: The low overhead (minor comparator/shifter logic) is negligible compared to the MAC datapath and memory, resulting in no expected change to frequency, throughput, or power consumption.**

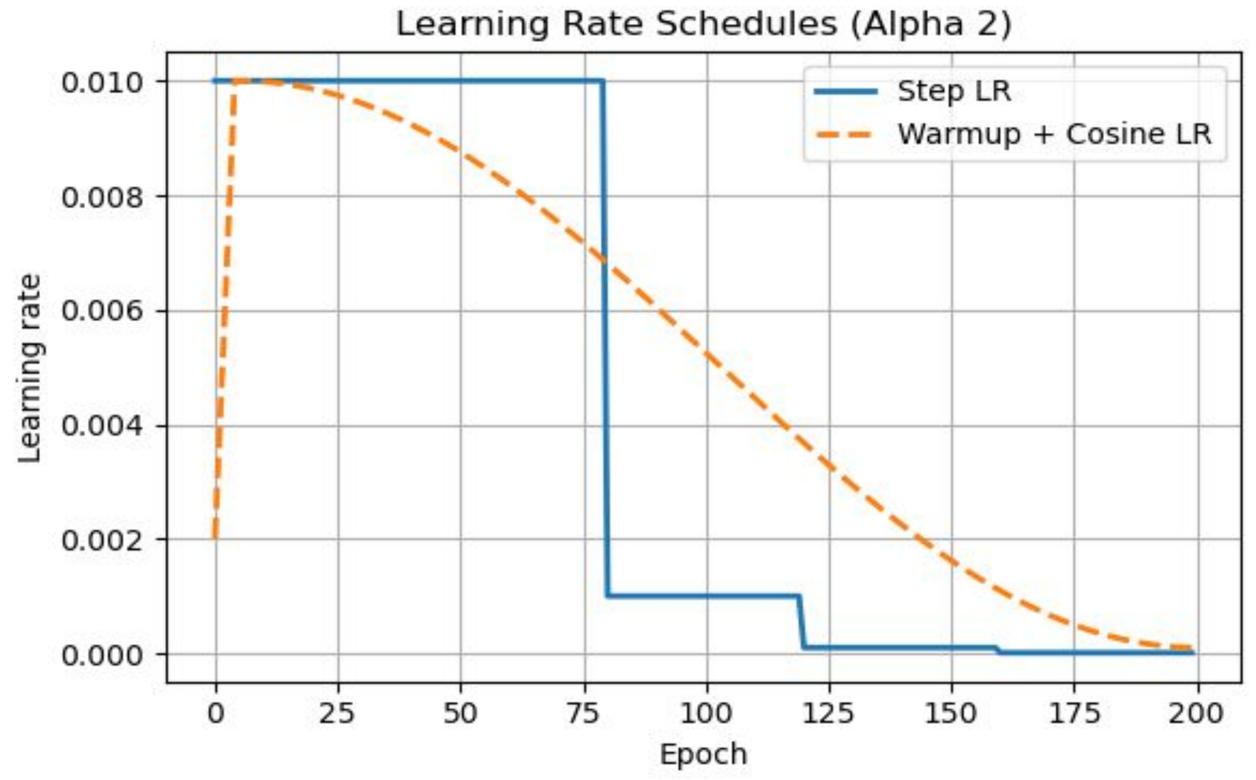


ReLU activation function

LeakyReLU activation function

Alpha 2 (Cosine Annealing Learning Rate)

- **Implementing Cosine Annealing with a 10-epoch warm-up improved convergence stability and overall accuracy by +0.82% on CIFAR-10.**
  - ○ **Hardware impact: Cosine annealing only changes the training schedule. The deployed fixed-point model and accelerator hardware remain identical, so throughput, power, and area are unchanged.**



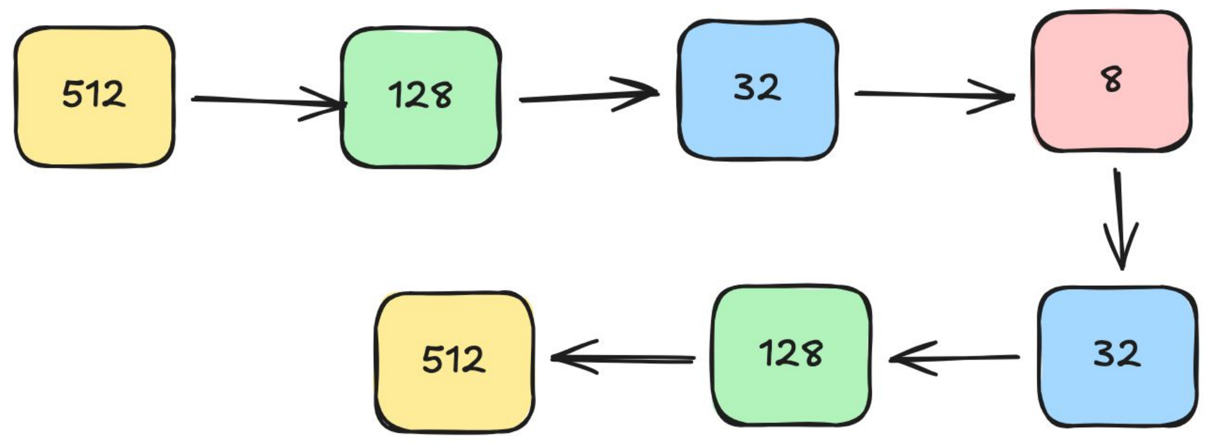
Alpha 3 (Activation Aware Pruning)

Sparsity	Pruning	Act-Aware	Hybrid
30%	90.37% accuracy	90.25% accuracy	90.11% accuracy
50%	90.45% accuracy	90.25% accuracy	89.25% accuracy
70%	90.22% accuracy	89.84% accuracy	86.85% accuracy

Metric	Pruning	Act-Aware	Improvement
TOPS/Watt	16.40	16.42	+0.1%
PE utilization	60.94%	60.94%	—
Zero Clustering	0.0405	0.0406	+0.4%

- **We used activation aware pruning to prune based on weight magnitude and average activation magnitudes.**
- **The accuracies were still comparable in performance after fine tuning and showed marginal improvement in the hardware efficiency**

Alpha 4 (Add Layers Before Bottleneck)



Added a gradual increase in the layers leading to the squeezed layer to improve accuracy