LatentAI

Adaptive AI for a Smarter Edge

Bit Efficient Quantization for Deep Neural Networks

Prateeth Nayak, David Zhang, Sek Chai*

*Latent AI and SRI International**

Quantization have afforded models that enable memory-efficient low-power inference. We present a comparison of data-free quantization schemes (i.e. asymmetric, symmetric, logarithmic) to explore limits below 8-bit precision. To better analyze quantization results, we describe the overall range and local sparsity of values afforded through various quantization schemes. We show the methods to lower bit-precision beyond quantization limits with object class clustering. We also highlight the connection of model architecture to quantization schemes.

Post-Training Model Quantization Schemes

Uniform Asymmetric

$$x_q = round\left[(x_f - min_{x_f})\frac{2^n - 1}{max_{x_f} - min_{x_f}}\right]$$

quantized tensor — *float-32 tensor* — *Bit-Precision = n*

Uniform Symmetric

$$x_q = round\left[(x_f)\frac{2^{n-1} - 1}{max|x_f|}\right]$$

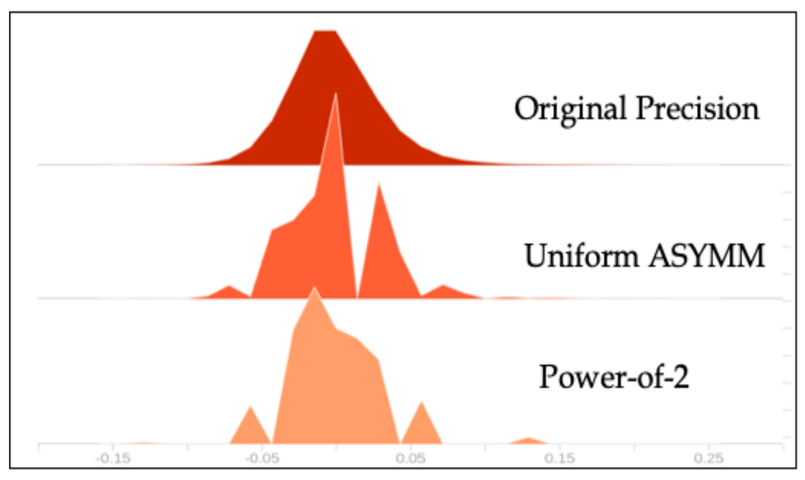
quantized tensor — *float-32 tensor* — *Bit-Precision = n*

Power Of Two

$$x_q = sign(x_f)2^{round(\log_2|x_f|)}$$

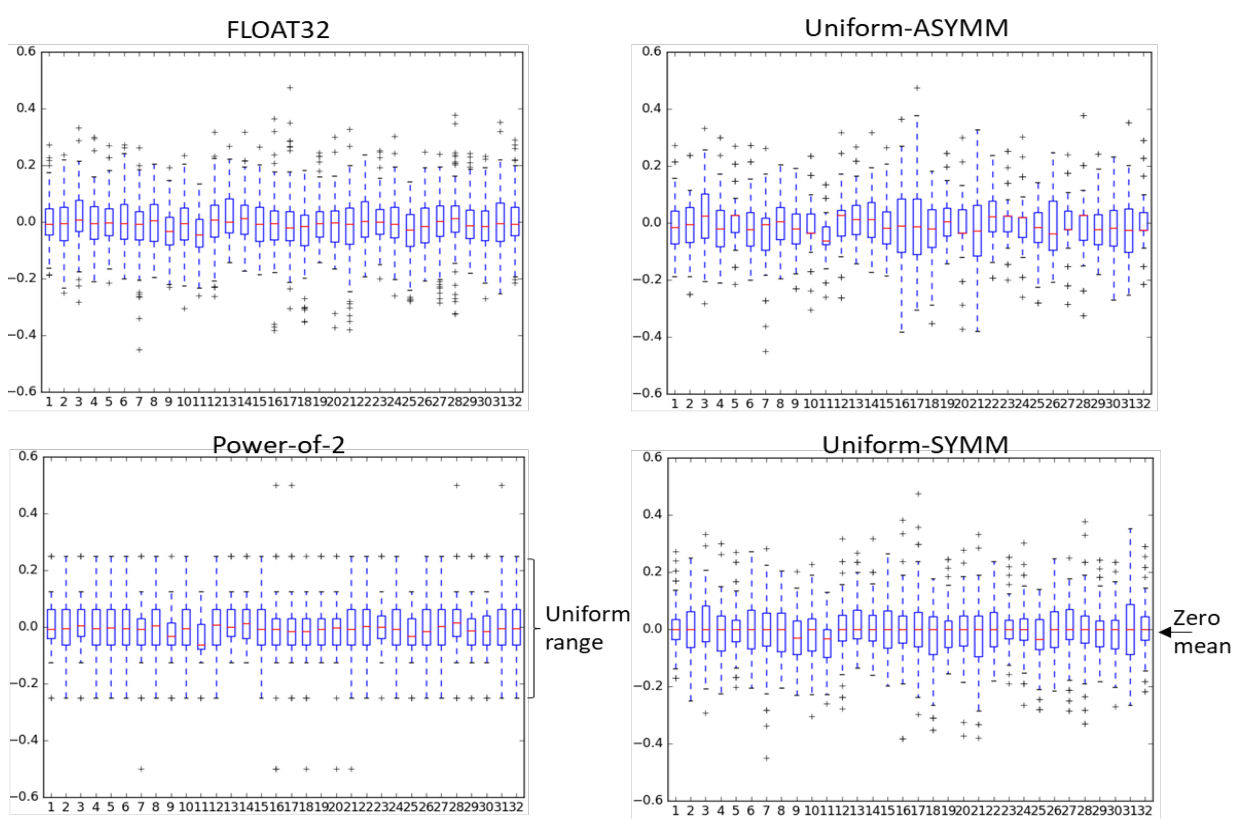
quantized tensor — *float-32 tensor*

The best quantization method is the one that is able to preserve the original model parameter distribution.



Histogram of weight distribution

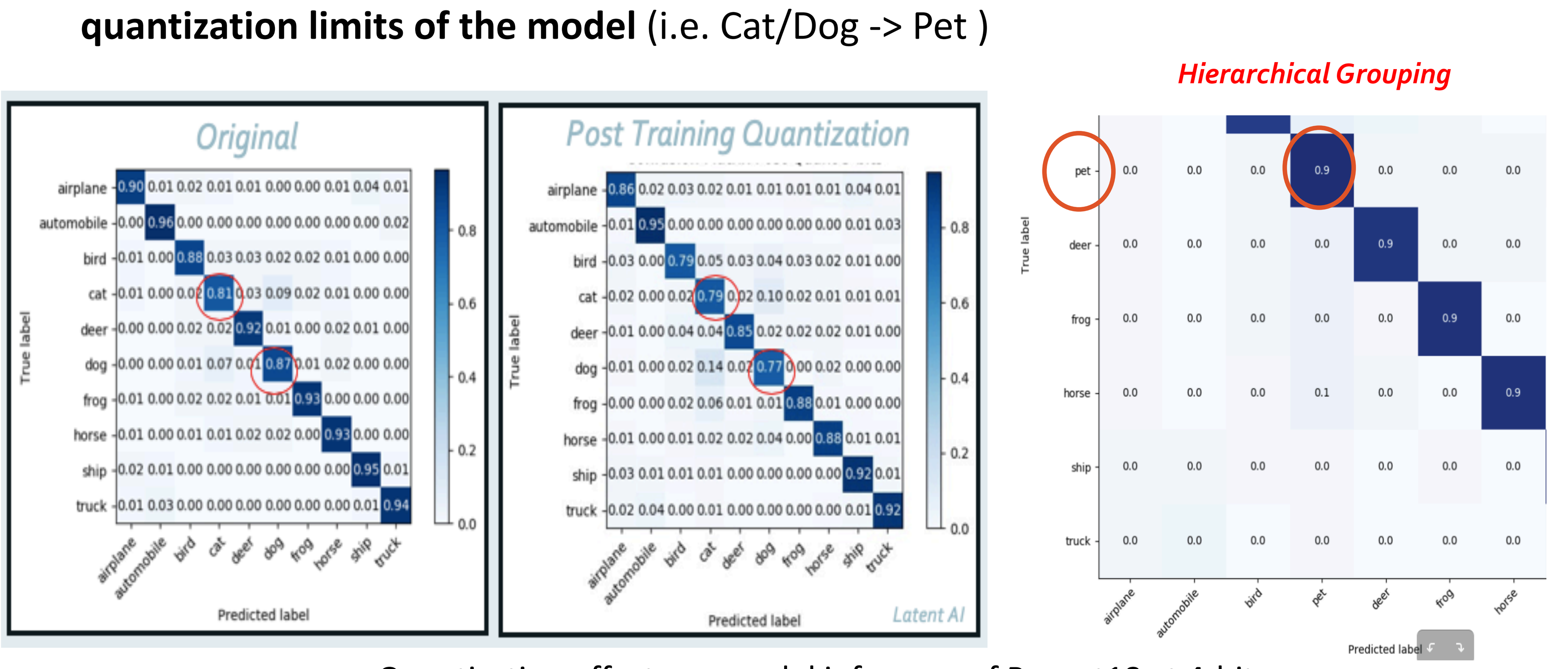
The uniform quantization approaches are able to preserve the mean of the baseline tensor, while the logarithmic approach of power-of-two maintains the range for each channel with lesser outlier parameters.



Quantization effects on quartile ranges of Tensors for different approaches. The FLOAT32 is the original tensor.

Hierarchical Clustering to Improve Accuracy

- **Quantization** to very low precision (4-bits) **creates biased-Inference**; Significance – Model parameters are quantization friendly if distributions of classes in training dataset are orthogonal in nature.
- Creating non-overlapping **Hierarchical Class-Distributions helps in pushing the quantization limits of the model** (i.e. Cat/Dog -> Pet)



Quantization effects on model inference of Resnet18 at 4-bits

Quantization Effect vs Model Architecture

Key Outcomes:

- Data-Free post-training **quantization achieves as low as 3-bit precision without affecting accuracy**.
- We can observe **40% reduction in model file size with 4% degradation in accuracy on Resnet18**, and **51% reduction in size with 1.65% drop in accuracy for SSD model** using partial quantization.
- The Table shows quantization results on Resnet18, Tiny-YOLOv2 and Mobilenet-SSD Models at prominent precision levels along with the file size compression results

	Acc	Size MB	Acc	Size MB	−Δ Acc	Δ Size	Acc	Size MB	−Δ Acc	Δ Size
Resnet18 CIFAR10 (Acc %)	90.87	2.00 (FP-32)	90.83	1.39	0.04% (8-bit)	30.5%	87.19	1.20	4% (4-bit)	40%
Tiny-YOLOv2 VOC 2007 (mAP)	52.97	58.8 (FP-32)	46.10	13.7	12.60% (8-bit)	76.7%	45.52	8.5	13.4% (6-bit)	85.5%
MobileNet-SSD Coco dataset (mAP)	28.56	63.28 (FP-32)	28.22	57.12	1.19% (Partial 8b)	9.73%	28.09	30.48	1.65% (Full backbone 8b)	51.8%

Post-Training Quantization results using Uniform ASYMM quantizer

	Acc	# bit	Size MB	Acc	# Avg-bit	Size MB	−Δ Acc	Δ bits
Resnet18 CIFAR10 (Acc %)	90.87	32	1.1	88.90	4~5	0.15	1.9%	90.15%
Tiny YOLOv2. VOC 2007 (mAP)	52.97	32	58.8	44.72	5~6	7.7	12.2%	83.7%

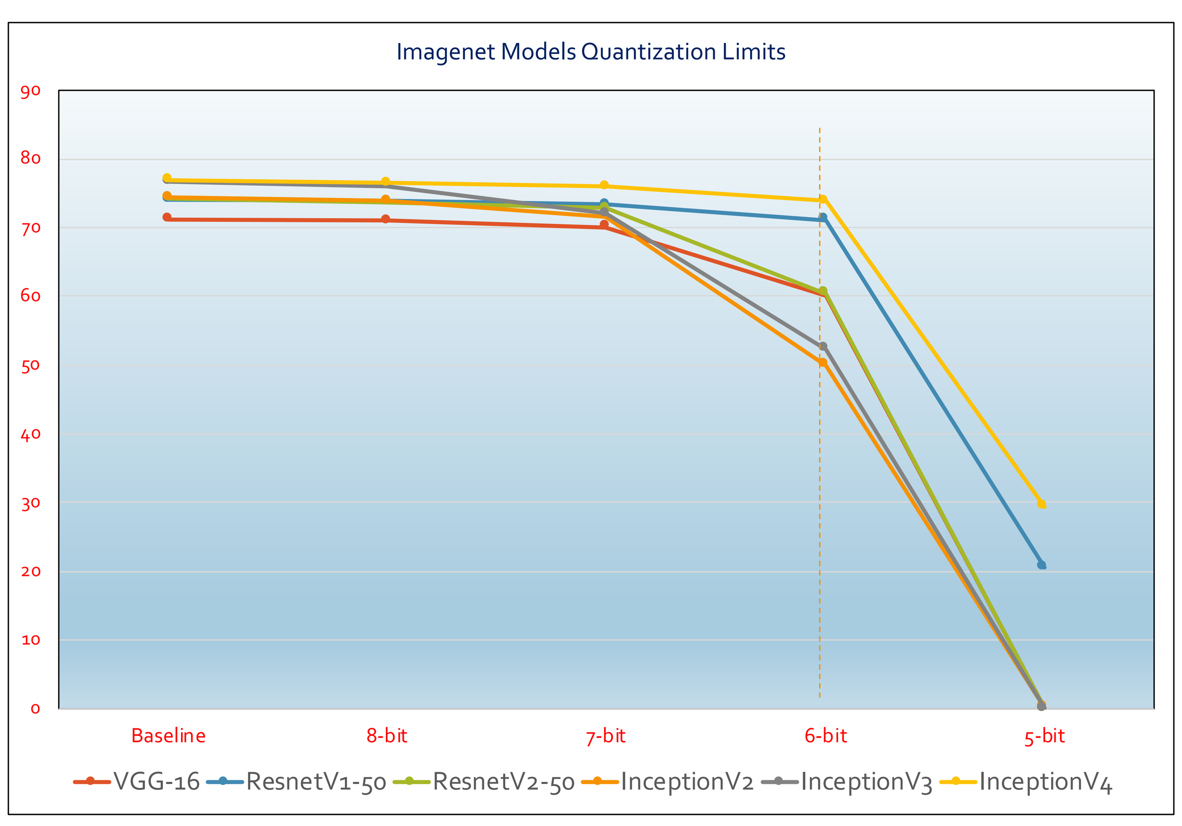
Training-Aware Quantization results using Power-of-2

Model Architecture	Baseline Model Size (MB)	Quantization Approach	Post-Quant Model Size (6-bit) (MB)		
			G-zip	7-zip	Δ Size
VGG-16	513.72	Uniform	95.78	69.65	27.2%
		Power-of-2	95.56	68.14	28.7%
ResnetV2-50	95.4	Uniform	14.05	10.12	28.0%
		Power-of-2	17.84	11.18	37.3%
InceptionV4	171.26	Uniform	26.6	19.4	27.1%
		Power-of-2	28.6	21.6	24.5%

Compression numbers using GZip and 7zip

Model Architecture Agnostic

- We also observe that the **quantization effect is model architecture agnostic**.
- It is more closely tied to the distribution of dataset the model has reached convergence optimality.
- Graph shows the quantization bit-limit for all the models that have converged on ImageNet Dataset. Accuracy vs Bit-Precision shows 6-bit limit.



Conclusion And Future Work

- **Quantized performance is closely tied to the dataset distribution**. For classification tasks, the **hierarchical grouping of overlapping class distribution gives lesser degradation on inference at lower bit precision**. For **regression tasks**, it is **still a challenge** to regress to coordinates with lesser precision.
- **Quantization effects can be independent of the model architecture**, e.g. for common feed forward convolution networks. **We observed that quantizing initial layers affects model performance the most**, suggesting (1) the **need to preserve initial learned features**, and (2) better **returns with quantization of semantic layers.**
- Using the framework built we are **able to deploy models to general purpose processors**, however **work still remains** in targeting **hardware constraints for optimized low-precision operations** for taking the full benefit of quantization schemes.

The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019