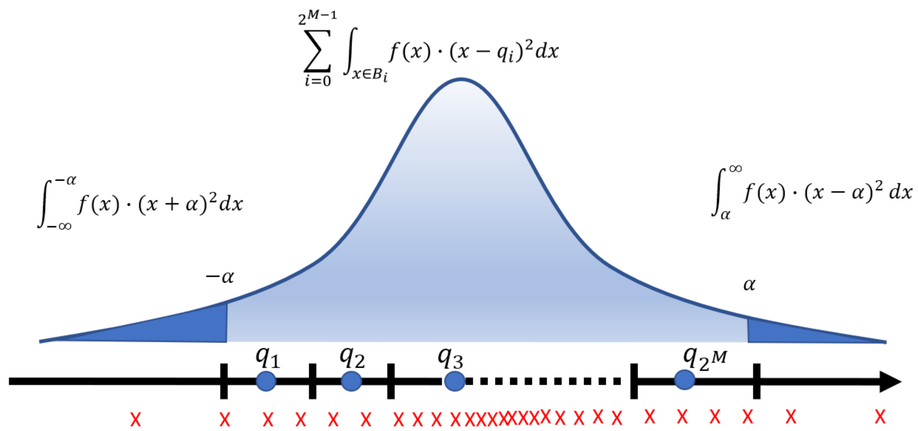
Post training 4-bit quantization of convolutional networks for rapid-deployment

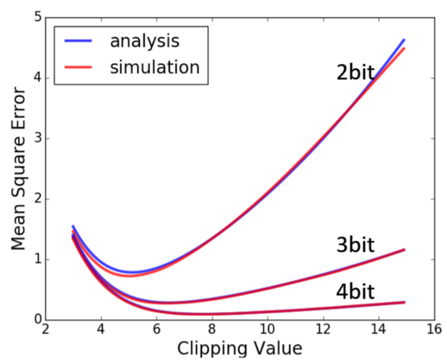
Ron Banner, Yury Nahshan and Daniel Soudry

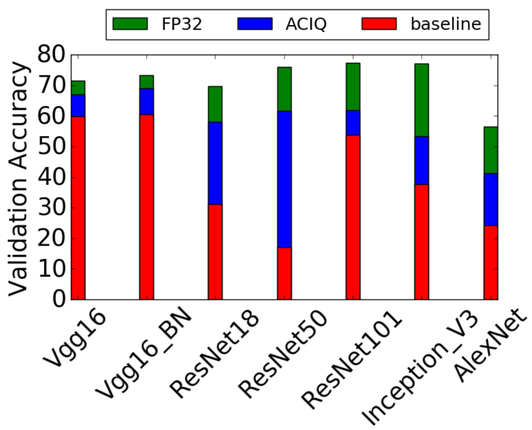
- 4-bit Post training quantization of weights and activations
  - No retraining
  - No data set

Method	VGG	VGG-BN	IncepV3	Res18	Res50	Res101
Baseline	67.2%	64.5%	30.6%	51.6%	62.0%	62.6%
All methods combined	70.5%	71.8%	66.4%	67.0%	73.8%	75.0%
Reference (FP32)	71.6%	73.4%	77.2%	69.7%	76.1%	77.3%

Analytical Clipping

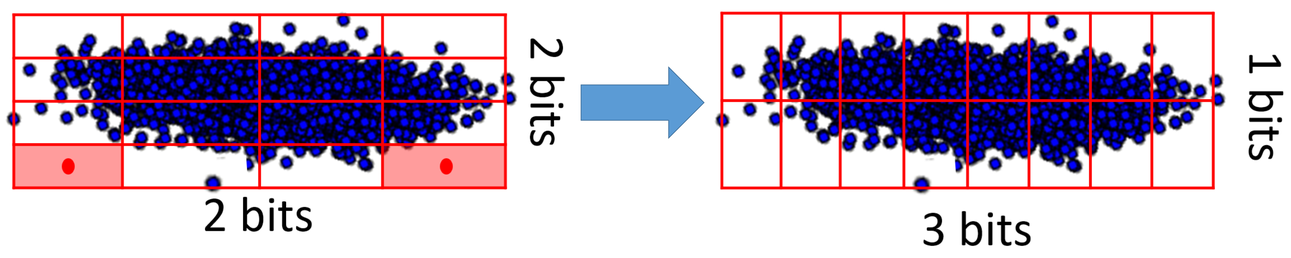


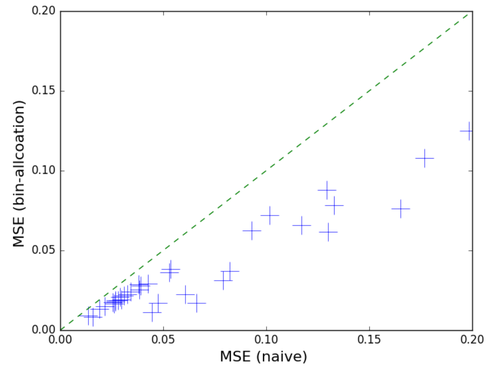


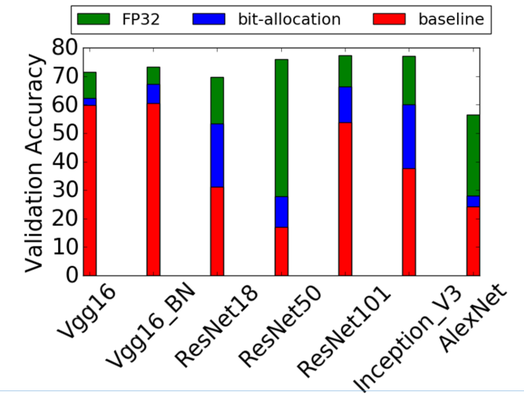


Bit-allocation

$$M_i = \left\lfloor \log_2 \left(\frac{\alpha_i^{\frac{2}{3}}}{\sum_i \alpha_i^{\frac{2}{3}}} \cdot B \right) \right\rfloor$$







Bias-correction

$$\mu_c = \mathbb{E}(W_c) - \mathbb{E}(W_c^q)$$

$$\xi_c = \frac{||W_c - \mathbb{E}(W_c)||_2}{||W_c^q - \mathbb{E}(W_c^q)||_2}$$

$$w \longleftarrow \xi_c (w + \mu_c), \quad \forall w \in W_c^q$$