Up or Down? Adaptive Rounding for Post-Training Quantization

PMLR, 2020

Qualcomm AI

INTRODUCTION

Rounding-to-Nearest

- Issues
  - When quantize NN, rounding-to-nearest is the predominant approach
  - This is not the best
- Objective
  - Provide a better weight rounding mechanism for post-training quantization

$$\mathbf{X}_{\text{quant}} = \text{round}\left(\text{scale} \cdot \mathbf{X} + \text{zeropoint}\right)$$

$$\mathbf{X}_{\text{dequant}} = \frac{\mathbf{X}_{\text{quant}} - \text{zeropoint}}{\text{scale}}$$

INTRODUCTION

Motivation: rounding-to-nearest is not optimal

- We want to minimize the task loss after quantization
- (a): second order Taylor series expansion (approximate $L(x, y, \mathrm{w} + \Delta w)$ using $L(x, y, w)$)

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2$$

$a = w$

$x = w + \Delta w$

After quantization

$$\mathbb{E}\left[\mathcal{L}\left(\mathbf{x}, \mathbf{y}, \mathbf{w} + \Delta\mathbf{w}\right) - \mathcal{L}\left(\mathbf{x}, \mathbf{y}, \mathbf{w}\right)\right] \quad (2)$$

$$\stackrel{(a)}{\approx} \mathbb{E}\left[\Delta\mathbf{w}^T \cdot \nabla_{\mathbf{w}}\mathcal{L}\left(\mathbf{x}, \mathbf{y}, \mathbf{w}\right) + \frac{1}{2}\Delta\mathbf{w}^T \cdot \nabla^2_{\mathbf{w}}\mathcal{L}\left(\mathbf{x}, \mathbf{y}, \mathbf{w}\right) \cdot \Delta\mathbf{w}\right] \quad (3)$$

$$= \Delta\mathbf{w}^T \cdot \mathbf{g}^{(\mathbf{w})} + \frac{1}{2}\Delta\mathbf{w}^T \cdot \mathbf{H}^{(\mathbf{w})} \cdot \Delta\mathbf{w}, \quad (4)$$

INTRODUCTION

Motivation: rounding-to-nearest is not optimal

- If the network is trained to convergence, gradient term will be close to 0

$$\Delta \mathbf{w}^T \cdot \mathbf{g}^{(\mathbf{w})} + \frac{1}{2} \Delta \mathbf{w}^T \cdot \mathbf{H}^{(\mathbf{w})} \cdot \Delta \mathbf{w},$$

If objective function using rounding-to-nearest is always the minimal value,

it is optimal method

INTRODUCTION

$$\Delta \mathbf{w}^T \cdot \mathbf{g}^{(\mathbf{w})} + \frac{1}{2}\Delta \mathbf{w}^T \cdot \mathbf{H}^{(\mathbf{w})} \cdot \Delta \mathbf{w},$$

Motivation: rounding-to-nearest is not optimal

- If the network is trained to convergence, gradient term will be close to 0

Example 1. Assume $\Delta \mathbf{w}^T = [\Delta w_1 \quad \Delta w_2]$ and

$$\mathbf{H}^{(\mathbf{w})} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix},$$

$$\Delta \mathbf{w}^T \cdot \mathbf{H}^{(\mathbf{w})} \cdot \Delta \mathbf{w} = \Delta \mathbf{w}_1^2 + \Delta \mathbf{w}_2^2 + \Delta \mathbf{w}_1 \Delta \mathbf{w}_2.$$

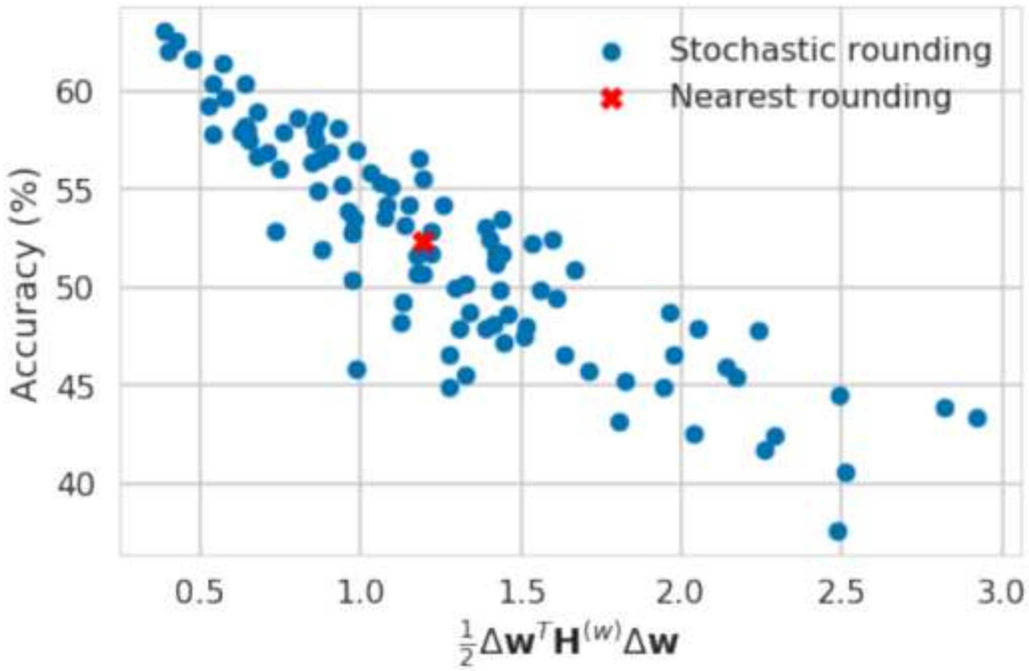
It is optimal considering the first two terms

INTRODUCTION

Motivation: rounding-to-nearest is not optimal

- ImageNet validation accuracy
- 4-bit quantization of the first layer of ResNet18

Rounding scheme	Acc(%)
Nearest	52.29
Ceil	0.10
Floor	0.10
Stochastic	52.06±5.52
Stochastic (best)	63.06



Stochastic quantization (100)

$$\mathrm{Int}(x) = \begin{cases} \lfloor x \rfloor & \text{with probability } \lceil x \rceil - x, \\ \lceil x \rceil & \text{with probability } x - \lfloor x \rfloor. \end{cases}$$

INTRODUCTION

Motivation: rounding-to-nearest is not optimal

- Set up two objective functions to solve the optimization problem: (13), (20)
- Re-design objective functions due to their complexity issues

$$\underset{\Delta\mathbf{w}^{(\ell)}}{\arg\min} \quad \mathbb{E}\left[\Delta\mathbf{w}^{(\ell)^T}\mathbf{H}^{(\mathbf{w}^{(\ell)})}\Delta\mathbf{w}^{(\ell)}\right]. \qquad (13)$$

$$= \underset{\Delta\mathbf{W}_{k,:}^{(\ell)}}{\arg\min} \quad \mathbb{E}\left[\left(\Delta\mathbf{W}_{k,:}^{(\ell)}\mathbf{x}^{(\ell-1)}\right)^2\right], \qquad (20)$$

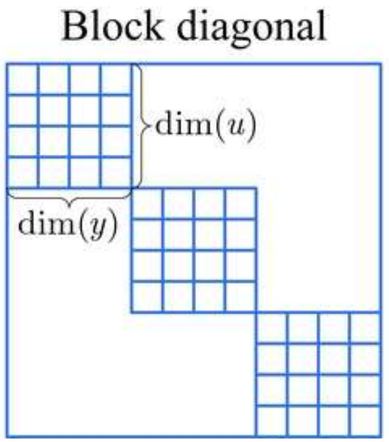
$$\underset{\mathbf{V}}{\arg\min} \quad \left\|\mathbf{W}\mathbf{x} - \widetilde{\mathbf{W}}\mathbf{x}\right\|_F^2 + \lambda f_{reg}(\mathbf{V}), \qquad (21)$$

INTRODUCTION

Objective Function 1.

- Assume gradient term will be close to 0
- Assume a block diagonal $H^{(W)}$
    - Ignore the interactions among weights belonging to different layers
    - Per-layer optimization problem

$$\underset{\Delta\mathbf{w}^{(\ell)}}{\arg\min} \quad \mathbb{E}\left[\Delta\mathbf{w}^{(\ell)^T}\mathbf{H}^{(\mathbf{w}^{(\ell)})}\Delta\mathbf{w}^{(\ell)}\right]. \qquad (13)$$



INTRODUCTION

$$\underset{\Delta \mathbf{w}^{(\ell)}}{\arg\min} \quad \mathbb{E}\left[\Delta \mathbf{w}^{(\ell)^T} \mathbf{H}^{(\mathbf{w}^{(\ell)})} \Delta \mathbf{w}^{(\ell)}\right]. \qquad (13)$$

Objective Function 2.

- (13) is an NP-hard optimization problem
- $H^{(W)}$ suffers from computational and complexity issues

For two weights in the same layer,

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{W}_{i,j}^{(\ell)} \partial \mathbf{W}_{m,o}^{(\ell)}} = \frac{\partial}{\partial \mathbf{W}_{m,o}^{(\ell)}} \left[\frac{\partial \mathcal{L}}{\partial \mathbf{z}_i^{(\ell)}} \cdot \mathbf{x}_j^{(\ell-1)}\right] = \frac{\partial^2 \mathcal{L}}{\partial \mathbf{z}_i^{(\ell)} \partial \mathbf{z}_m^{(\ell)}} \cdot \mathbf{x}_j^{(\ell-1)} \mathbf{x}_o^{(\ell-1)},$$

Complexity issues are mainly caused by

$$\mathbf{H}^{(\mathbf{w}^{(\ell)})} = \mathbb{E}\left[\mathbf{x}^{(\ell-1)} \mathbf{x}^{(\ell-1)^T} \otimes \nabla^2_{\mathbf{z}^{(\ell)}} \mathcal{L}\right],$$

$z_i^{(l)}$: pre-activation for layer l

INTRODUCTION

$$\underset{\Delta \mathbf{w}^{(\ell)}}{\arg\min} \quad \mathbb{E}\left[\Delta \mathbf{w}^{(\ell)^T} \mathbf{H}^{(\mathbf{w}^{(\ell)})} \Delta \mathbf{w}^{(\ell)}\right]. \qquad (13)$$

$$\mathbf{H}^{(\mathbf{w}^{(\ell)})} = \mathbb{E}\left[\mathbf{x}^{(\ell-1)} \mathbf{x}^{(\ell-1)^T} \otimes \nabla^2_{\mathbf{z}^{(\ell)}} \mathcal{L}\right],$$

Objective Function 2.

- Assume gradient term will be close to 0
- Assume $\nabla^2_{z^{(l)}} L$ is a diagonal matrix
- Assume $\nabla^2_{z^{(l)}} L_{i,i}$ is a constant independent of the input data samples (strong assumption)

$$\underset{\Delta \mathbf{W}^{(\ell)}_{k,:}}{\arg\min} \quad \mathbb{E}\left[\nabla^2_{\mathbf{z}^{(\ell)}} \mathcal{L}_{k,k} \cdot \Delta \mathbf{W}^{(\ell)}_{k,:} \mathbf{x}^{(\ell-1)} \mathbf{x}^{(\ell-1)^T} \Delta \mathbf{W}^{(\ell)^T}_{k,:}\right] \qquad (18)$$

$$\stackrel{(a)}{=} \underset{\Delta \mathbf{W}^{(\ell)}_{k,:}}{\arg\min} \quad \Delta \mathbf{W}^{(\ell)}_{k,:} \mathbb{E}\left[\mathbf{x}^{(\ell-1)} \mathbf{x}^{(\ell-1)^T}\right] \Delta \mathbf{W}^{(\ell)^T}_{k,:} \qquad (19)$$

$$= \underset{\Delta \mathbf{W}^{(\ell)}_{k,:}}{\arg\min} \quad \mathbb{E}\left[\left(\Delta \mathbf{W}^{(\ell)}_{k,:} \mathbf{x}^{(\ell-1)}\right)^2\right], \qquad (20)$$

$z_i^{(l)}$: pre-activation for layer l

METHOD

$$\mathbb{E}\left[\mathcal{L}\left(\mathbf{x},\mathbf{y},\mathbf{w}+\Delta\mathbf{w}\right)-\mathcal{L}\left(\mathbf{x},\mathbf{y},\mathbf{w}\right)\right]$$

AdaRound

- s: fixed scaling factor
- $f_{reg}(.)$: regularizer
- $h(.)$: $\in [0,1]$ rectified sigmoid

$$\underset{\mathbf{V}}{\arg\min} \quad \left\|\mathbf{W}\mathbf{x}-\widetilde{\mathbf{W}}\mathbf{x}\right\|_F^2+\lambda f_{reg}\left(\mathbf{V}\right), \qquad (21)$$

$$\widetilde{\mathbf{W}}=\mathrm{s}\cdot clip\left(\left\lfloor\frac{\mathbf{W}}{\mathrm{s}}\right\rfloor+h\left(\mathbf{V}\right),\mathrm{n},\mathrm{p}\right).$$

METHOD

$$\mathbb{E}\left[\mathcal{L}\left(\mathbf{x}, \mathbf{y}, \mathbf{w}+\Delta \mathbf{w}\right)-\mathcal{L}\left(\mathbf{x}, \mathbf{y}, \mathbf{w}\right)\right]$$

AdaRound

$$\underset{\mathbf{V}}{\arg\min} \quad \left\|\mathbf{W}\mathbf{x}-\widetilde{\mathbf{W}}\mathbf{x}\right\|_F^2+\lambda f_{reg}\left(\mathbf{V}\right), \qquad (21)$$

$$\widetilde{\mathbf{W}}=\mathrm{s}\cdot clip\left(\left\lfloor\frac{\mathbf{W}}{\mathrm{s}}\right\rfloor+h\left(\mathbf{V}\right), \mathrm{n}, \mathrm{p}\right).$$

$$f_{reg}\left(\mathbf{V}\right)=\sum_{i,j} 1-\left|2h\left(\mathbf{V}_{i,j}\right)-1\right|^{\beta},$$

1.1 -0.1

$$h\left(\mathbf{V}_{i,j}\right)=clip(\sigma\left(\mathbf{V}_{i,j}\right)\left(\zeta-\gamma\right)+\gamma, 0, 1),$$

Rectified sigmoid has non-vanishing gradients as $h(V)$ approaches 0 or 1

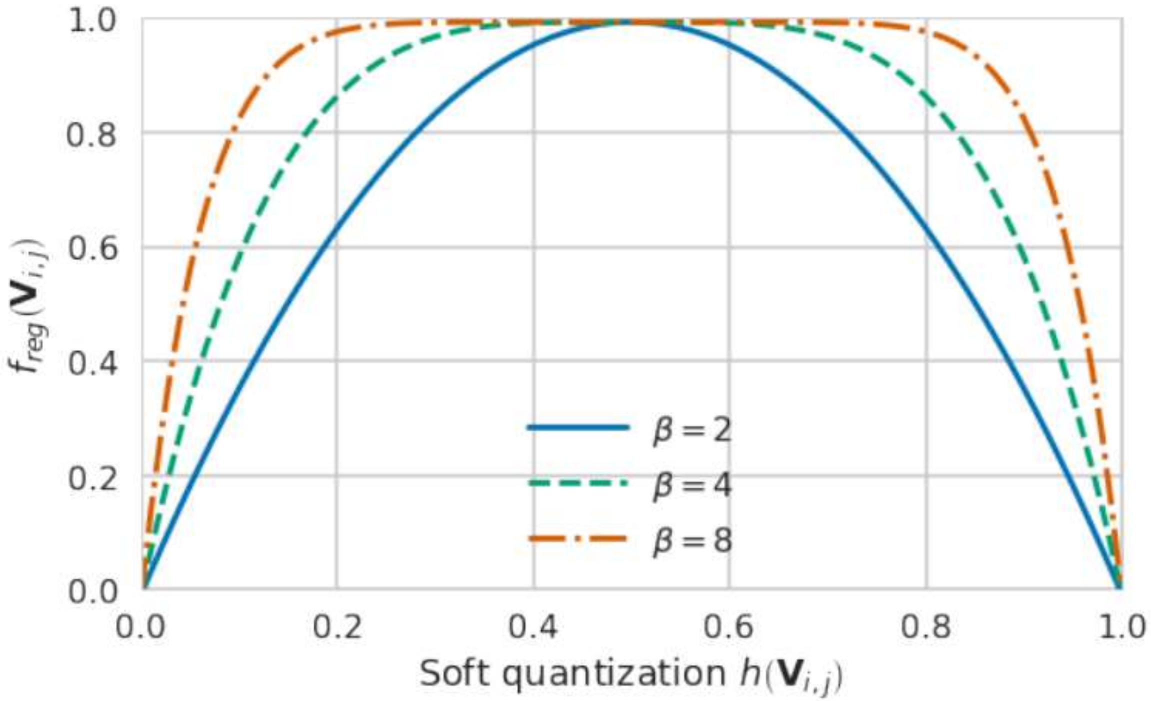
METHOD

AdaRound

- Initial phase: higher β
- Layer phase: lower β

$$f_{reg}(\mathbf{V}) = \sum_{i,j} 1 - |2h(\mathbf{V}_{i,j}) - 1|^{\beta},$$

$$h(\mathbf{V}_{i,j}) = clip(\sigma(\mathbf{V}_{i,j})(\overset{1.1}{\zeta} - \overset{-0.1}{\gamma}) + \gamma, 0, 1),$$



METHOD

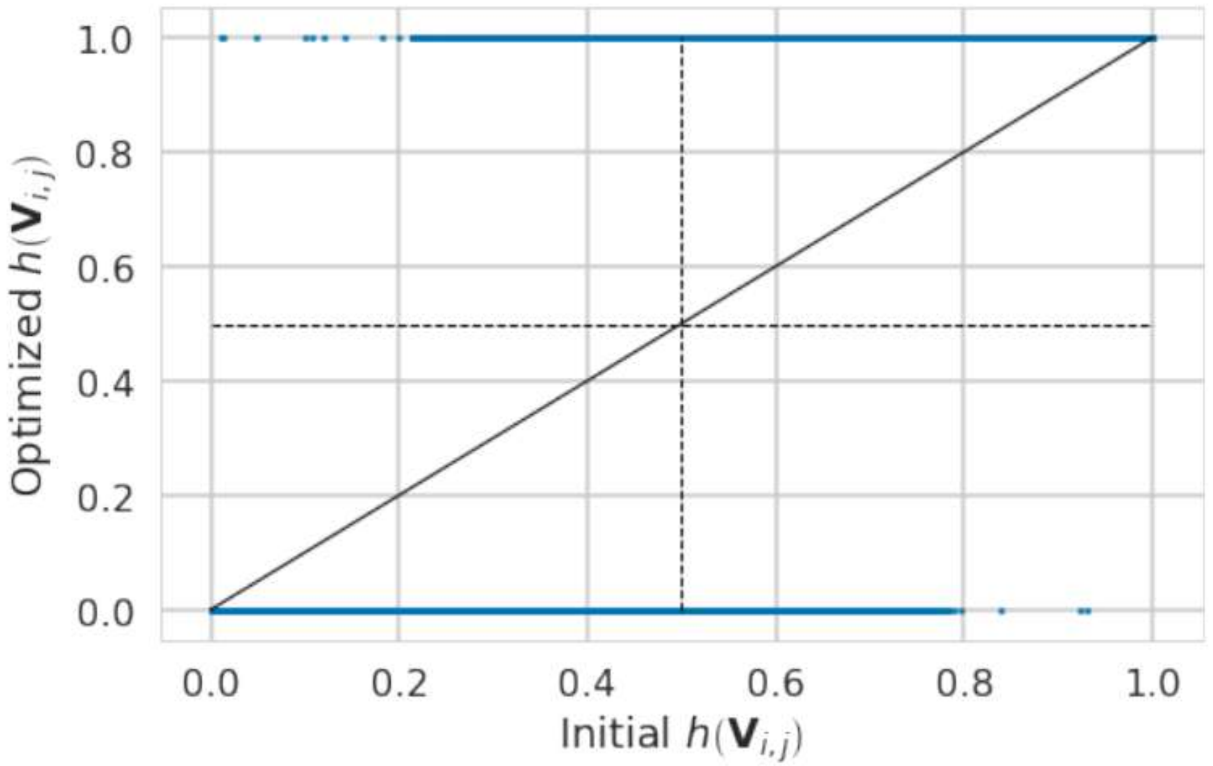
AdaRound

- Initial phase: higher β
- Layer phase: lower β

$$f_{reg}(\mathbf{V}) = \sum_{i,j} 1 - |2h(\mathbf{V}_{i,j}) - 1|^{\beta},$$

1.1 -0.1

$$h(\mathbf{V}_{i,j}) = clip(\sigma(\mathbf{V}_{i,j})(\zeta - \gamma) + \gamma, 0, 1),$$



METHOD

AdaRound

- (21) does not account for the quantization error introduced due to the previous layer
- In order to avoid the accumulation of quantization error for deeper network, use following formulation

$$\underset{\mathbf{V}}{\arg\min} \quad \left\| \mathbf{W}\mathbf{x} - \widetilde{\mathbf{W}}\mathbf{x} \right\|_F^2 + \lambda f_{reg}(\mathbf{V}), \qquad (21)$$

$$\downarrow$$

$$\underset{\mathbf{V}}{\arg\min} \left\| f_a(\mathbf{W}\mathbf{x}) - f_a\left(\widetilde{\mathbf{W}}\hat{\mathbf{x}}\right) \right\|_F^2 + \lambda f_{reg}(\mathbf{V}), \quad (25)$$

EXPERIMENTS

Settings

- Symmetric 4-bit weight quantization
- Layer-wise quantization
- Use pre-defined scaling factor $||\mathbf{W} - \overline{\mathbf{W}}||_F^2$
- ResNet-18 (validation accuracy: 69.68%): Nvidia GTX 1080 Ti single GPU – 10 minutes
- 1024 unlabeled image on ImageNet training dataset
- Batch size: 32, Iteration: 10,000

EXPERIMENTS

Ablation study

- From task loss to local loss

$$\underset{\Delta\mathbf{w}^{(\ell)}}{\arg\min} \quad \mathbb{E}\left[\Delta\mathbf{w}^{(\ell)^T}\mathbf{H}^{(\mathbf{w}^{(\ell)})}\Delta\mathbf{w}^{(\ell)}\right]. \qquad (13)$$

$$= \underset{\Delta\mathbf{W}_{k,:}^{(\ell)}}{\arg\min} \quad \mathbb{E}\left[\left(\Delta\mathbf{W}_{k,:}^{(\ell)}\mathbf{x}^{(\ell-1)}\right)^2\right], \qquad (20)$$

$$\underset{\mathbf{V}}{\arg\min} \quad \left\|\mathbf{W}\mathbf{x} - \widetilde{\mathbf{W}}\mathbf{x}\right\|_F^2 + \lambda f_{reg}(\mathbf{V}), \qquad (21)$$

Rounding	First layer	All layers
Nearest	52.29	23.99
$\mathbf{H}^{(\mathbf{w})}$ task loss (cf. (13))	68.62±0.17	N/A
Local MSE loss (cf. (20))	69.39±0.04	65.83±0.14
Cont. relaxation (cf (21))	**69.58±0.03**	**66.56±0.12**

EXPERIMENTS

Ablation study

- Design choices for AdaRound

Rounding	First layer	All layers
Sigmoid + T annealing	69.31±0.21	65.22±0.67
Sigmoid + f_{reg}	**69.58±0.03**	66.25±0.15
Rect. sigmoid + f_{reg}	**69.58±0.03**	**66.56±0.12**

EXPERIMENTS

Ablation study

$$\underset{\mathbf{V}}{\arg\min} \quad \left\|\mathbf{W}\mathbf{x} - \widetilde{\mathbf{W}}\mathbf{x}\right\|_F^2 + \lambda f_{reg}\left(\mathbf{V}\right), \qquad (21)$$

$$\underset{\mathbf{V}}{\arg\min}\left\|f_a\left(\mathbf{W}\mathbf{x}\right) - f_a\left(\widetilde{\mathbf{W}}\hat{\mathbf{x}}\right)\right\|_F^2 + \lambda f_{reg}\left(\mathbf{V}\right), \quad (25)$$

Optimization	Acc (%)
Layer wise	66.56±0.12
Asymmetric	68.37±0.07
Asymmetric + ReLU	**68.60±0.09**

EXPERIMENTS

Ablation study

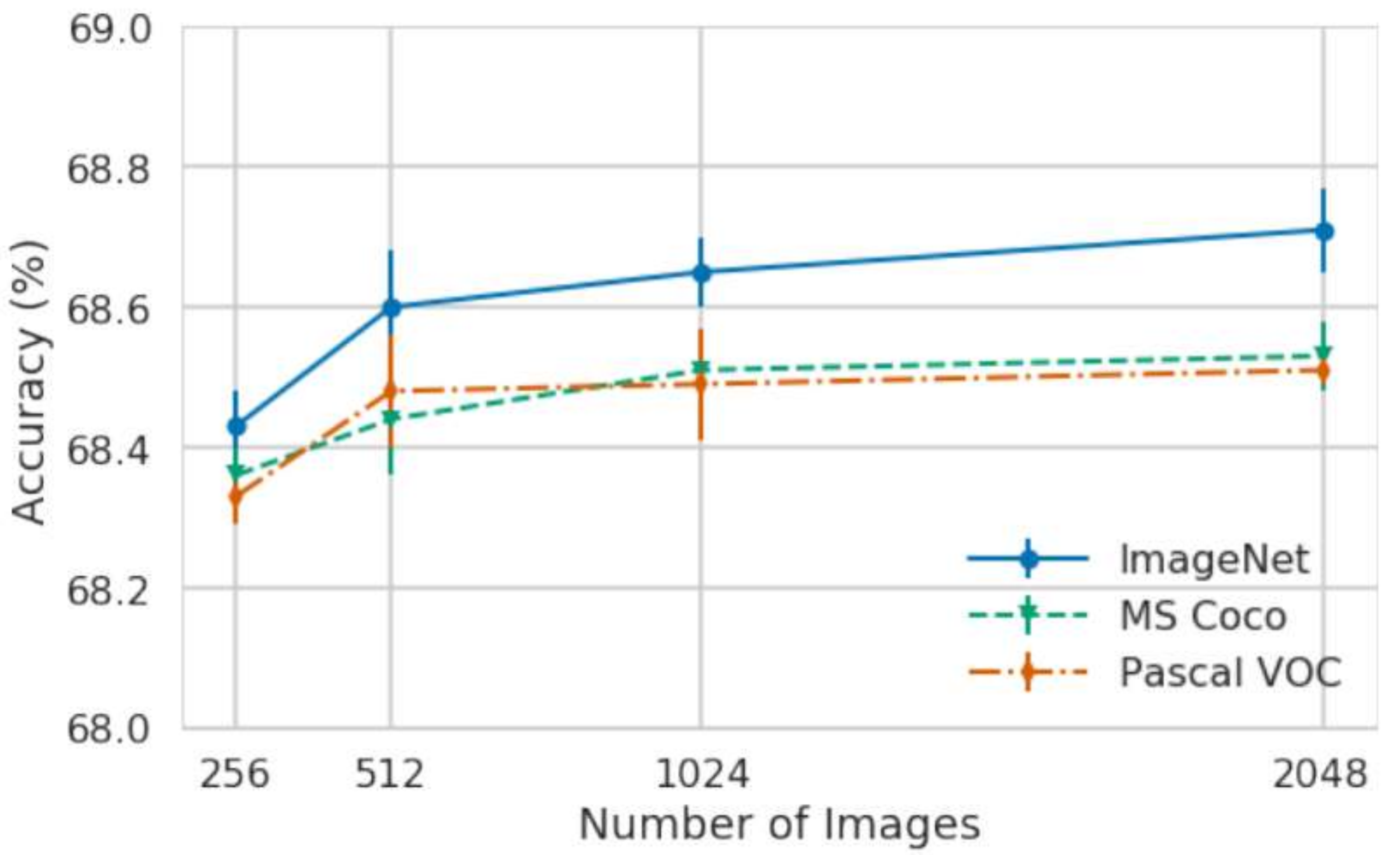
- Influence of quantization grid

Grid	Nearest	AdaRound
Min-Max	0.23	61.96±0.04
$\left\|\mathbf{W} - \overline{\mathbf{W}}\right\|_F^2$	23.99	**68.60±0.09**
$\left\|\mathbf{W}\mathbf{x} - \overline{\mathbf{W}}\widehat{\mathbf{x}}\right\|_F^2$	42.89	**68.62±0.08**

EXPERIMENTS

Ablation study

- Optimization robustness to data



EXPERIMENTS

Ablation study

- Different post-training quantization: channel-wise, do not quantize the first and the last layer
- AdaRound: layer-wise, 2048 samples, 20,000 iterations, activation quantization with Min-Max

Optimization	#bits W/A	Resnet18	Resnet50	InceptionV3	MobilenetV2
Full precision	32/32	69.68	76.07	77.40	71.72
DFQ (Nagel et al., 2019)	8/8	69.7	-	-	71.2
Nearest	4/32	23.99	35.60	1.67	8.09
OMSE+opt(Choukroun et al., 2019)	4*/32	67.12	74.67	73.66	-
OCS (Zhao et al., 2019)	4/32	-	66.2	4.8	-
AdaRound	4/32	**68.71±0.06**	**75.23±0.04**	**75.76±0.09**	**69.78±0.05**†
DFQ (our impl.)	4/8	38.98	52.84	-	46.57
Bias corr (Banner et al., 2019)	4*/8	67.4	74.8	59.5	-
AdaRound w/ act quant	4/8	**68.55±0.01**	**75.01±0.05**	**75.72±0.09**	**69.25±0.06**†

EXPERIMENTS

Ablation study

- Different post-training quantization: channel-wise, do not quantize the first and the last layer
- AdaRound: layer-wise, 2048 samples, 20,000 iterations, activation quantization with Min-Max

Optimization	#bits W/A	Resnet18	Resnet50	InceptionV3	MobilenetV2
Full precision	32/32	69.68	76.07	77.40	71.72
DFQ (Nagel et al., 2019)	8/8	69.7	-	-	71.2
Nearest	4/32	23.99	35.60	1.67	8.09
OMSE+opt(Choukroun et al., 2019)	4*/32	67.12	74.67	73.66	-
OCS (Zhao et al., 2019)	4/32	-	66.2	4.8	-
AdaRound	4/32	**68.71±0.06**	**75.23±0.04**	**75.76±0.09**	**69.78±0.05**†
DFQ (our impl.)	4/8	38.98	52.84	-	46.57
Bias corr (Banner et al., 2019)	4*/8	67.4	74.8	59.5	-
AdaRound w/ act quant	4/8	**68.55±0.01**	**75.01±0.05**	**75.72±0.09**	**69.25±0.06**†