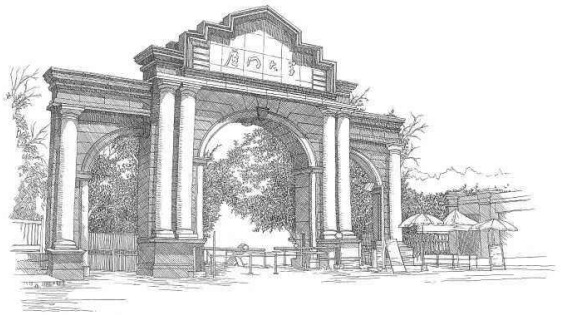
Solving Oscillation Problem in Post-Training Quantization Through a Theoretical Perspective

Yuexiao Ma, Huixia Li, Xiawu Zheng, Xuefeng Xiao,
Rui Wang, Shilei Wen, Xin Pan, Fei Chao, Rongrong Ji

TUE-PM-366

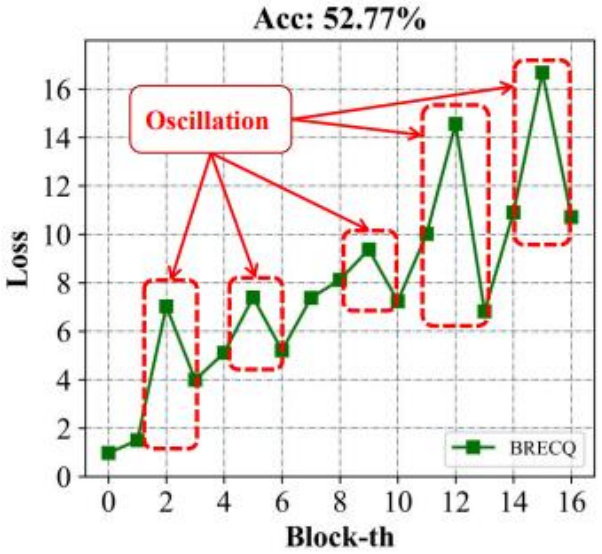


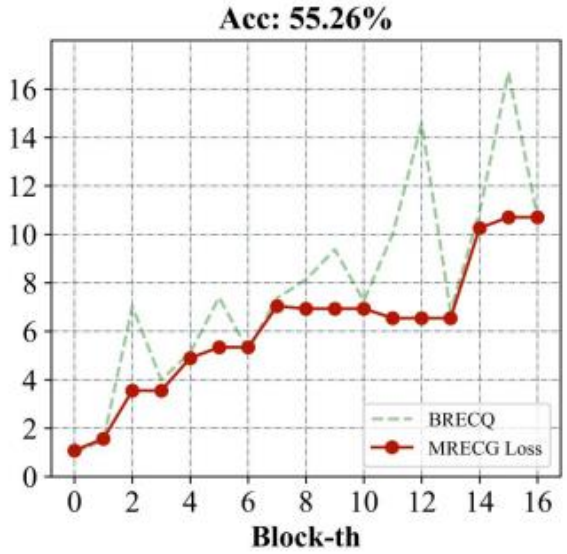
Solving Oscillation Problem in Post-Training Quantization Through a Theoretical Perspective

Contribution: **The PTQ oscillation problem is revealed for the first time** → This oscillation is caused by differences in the capacities of adjacent modules → We propose a novel Mixed REConstruction Granularity (MRECG) method to solve the oscillation problem → We validate the effectiveness of the proposed method in ImageNet

Analysis 1: The equivalence module will make the reconstruction loss incremental due to the cumulative effect of quantization error

Analysis 2: The oscillation of the reconstructed loss in post-training quantization is caused by the difference in capacity of adjacent modules





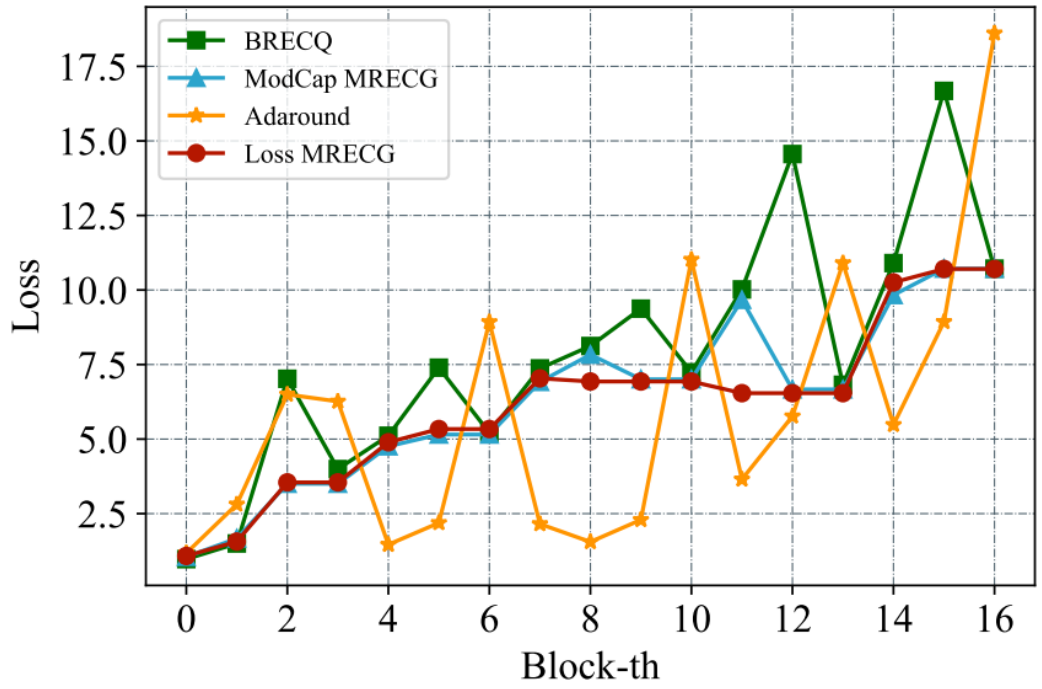
Traditional PTQ algorithm : Occurrence of reconstructed loss oscillation

MRECG: Smoothing out loss oscillations while bringing accuracy gains

The logical/correlation chain

Accuracy ∝ Final error ∝ The largest error ∝ Oscillation

Oscillation ∝ The largest error



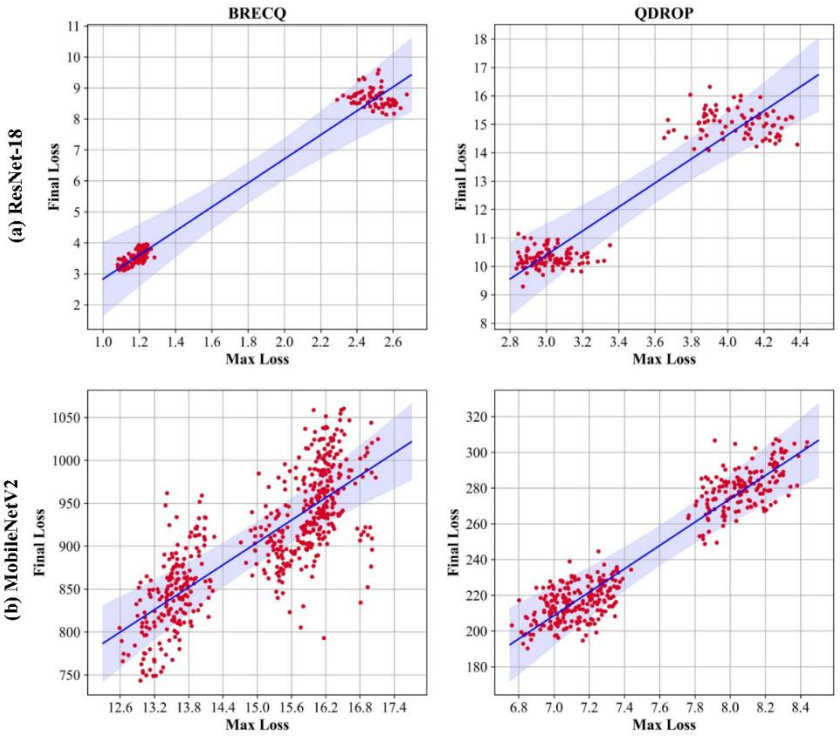
the more severe the oscillation, the larger the peak of the loss

The reconstruction loss distributions of different algorithms for 4/4 bit configurations on 0.5 scaled MobileNetV2, including Adaround, BRECQ, and MRECG.

The logical/correlation chain

Accuracy ∝ Final error ∝ The largest error ∝ Oscillation

Final error ∝ The largest error



the largest error is positively correlated with the final error.

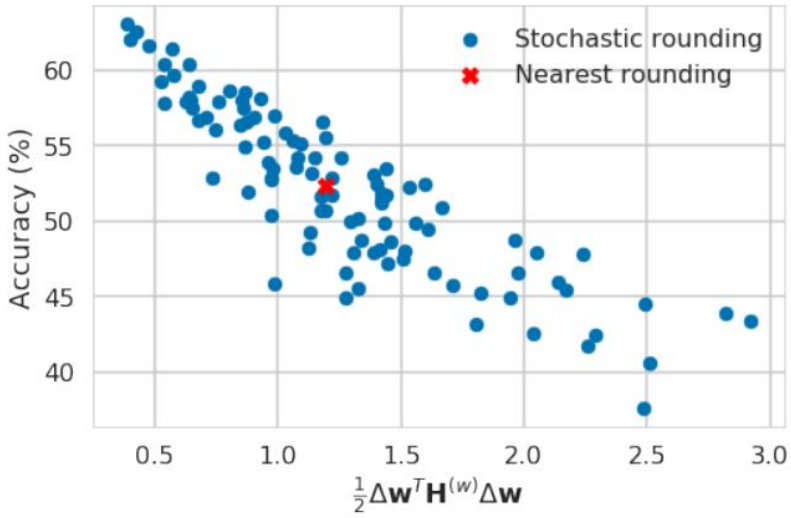
The relationship between the final reconstruction error and the maximum reconstruction error of the previous module. We randomly sample a number of mixed reconstruction granularity schemes and recover the accuracy using BRECQ or QDROP.

The logical/correlation chain

Accuracy ∝ Final error ∝ The largest error ∝ Oscillation

Accuracy ∝ Final error

Empirically:



Theoretically:

$$\mathbb{E}[L(\mathbf{w}+\Delta\mathbf{w})]-\mathbb{E}[L(\mathbf{w})] \approx \Delta\mathbf{w}^{\mathsf{T}}\bar{\mathbf{g}}^{(\mathbf{w})}+\frac{1}{2}\Delta\mathbf{w}^{\mathsf{T}}\bar{\mathbf{H}}^{(\mathbf{w})}\Delta\mathbf{w},$$
$$=\frac{1}{2}\Delta\mathbf{w}^{\mathsf{T}}\bar{\mathbf{H}}^{(\mathbf{w})}\Delta\mathbf{w}$$

BRECQ [1] and Adaround [2] conduct extensive experiments and perform some derivations to prove theoretically and empirically that the final error is positively correlated with accuracy.

Reason for oscillation and impact of calibration data batch size

Reason for oscillation：

Theorem 1. *Given a pre-trained model and input data. If two adjacent modules are equivalent, we have,*

$$\mathcal{L}(W_i, X_i) \leq \mathcal{L}(W_{i+1}, X_{i+1}). \tag{2}$$

Corollary 1. *Suppose two adjacent modules be topologically homogeneous. If the module capacity of the later module is large enough, the loss will decrease. Conversely, if the latter module capacity is smaller than the former, then the accumulation effect of quantization error is exacerbated.*

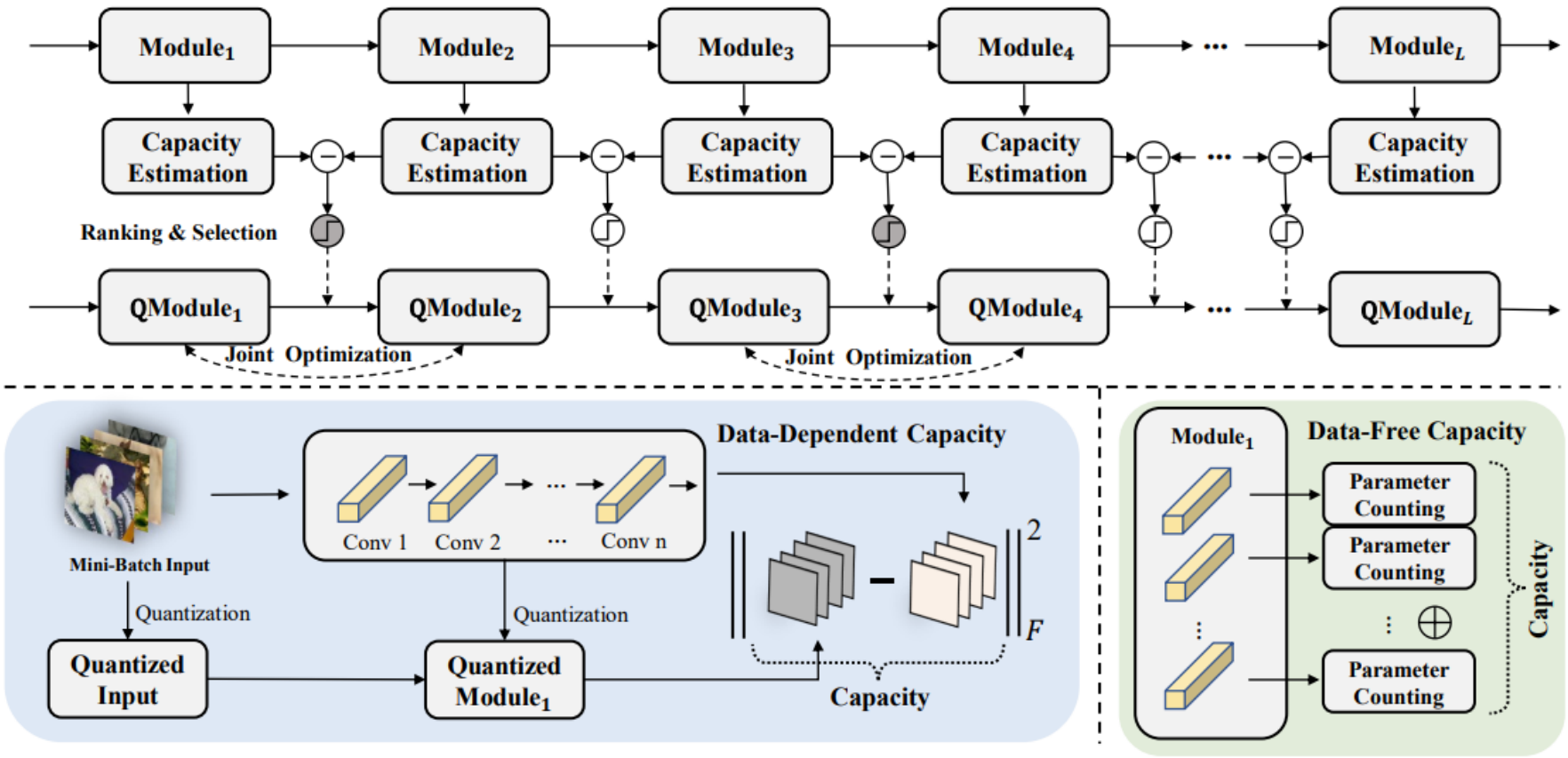
From the left theorem and corollary, the oscillation problem of PTQ on a wide range of models is caused by the excessive difference in capacities with adjacent modules.

Impact of calibration data batch size:

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{m=1}^{N} \| f_i^{(n)}(W_i, X_i^{(m)}) - f_i^{(n)}(\widetilde{W}_i, \widetilde{X}_i^{(m)}) \|_F^2$$
$$= \mathbb{E}\left[\| f_i^{(n)}(W_i, X_i) - f_i^{(n)}(\widetilde{W}_i, \widetilde{X}_i) \|_F^2 \right],$$

The law of large numbers proves that the mean of samples is infinitely close to the expectation when the sample size N tends to infinity. In other words, expanding the batch size of calibration data reduces the expected approximation error, thus benefiting accuracy

Optimized Solutions



$$\underset{\mathbf{m}}{\operatorname{argmin}} \sum_{l=1}^{L-1} (CM_l - CM_{l+1})^2 m_l + \lambda(\mathbf{m} \cdot \mathbb{1} - k)^2$$

Smoothing out loss oscillations by **jointly optimizing** neighboring modules with large differences in module capabilities

Experiments

Methods	W/A	Res18	Res50	MBV2×1.0	MBV2×0.75	MBV2×0.5	MBV2×0.35
Full Prec.	32/32	71.01	76.63	72.20	69.95	64.60	60.08
ACIQ-Mix [1]	4/4	67.00	73.80	-	-	-	-
ZeroQ [2]	4/4	21.71	2.94	26.24	-	-	-
LAPQ [22]	4/4	60.30	70.00	49.70	-	-	-
AdaQuant [10]	4/4	69.60	75.90	47.16	-	-	-
Bit-Split [29]	4/4	67.56	73.71	-	-	-	-
AdaRound [21]	4/4	67.96	73.88	61.52	55.32	40.71	35.13
BRECQ* [16]	4/4	68.69	74.88	67.51	62.94	53.02	48.88
Ours+BRECQ	4/4	**69.06** (+0.37)	74.84	**68.56** (+1.05)	**64.55** (+1.61)	**55.26** (+2.24)	**50.67** (+1.79)
QDROP [30]	4/4	69.10	75.03	67.89	63.26	54.19	49.79
Ours+QDROP	4/4	**69.46** (+0.36)	**75.35** (+0.32)	**68.84** (+0.95)	**64.39** (+1.13)	**55.64** (+1.45)	**50.94** (+1.15)
LAPQ [22]	2/4	0.18	0.14	0.13	-	-	-
AdaQuant [10]	2/4	0.11	0.12	0.15	-	-	-
AdaRound [21]	2/4	62.12	66.11	36.31	25.58	15.12	12.46
BRECQ* [16]	2/4	63.71	68.55	52.30	47.14	34.55	30.80
Ours+BRECQ	2/4	**65.61** (+1.9)	**70.04** (+1.49)	**58.49** (+6.19)	**52.50** (+5.36)	**41.16** (+6.61)	**35.46** (+4.66)
QDROP [30]	2/4	64.66	70.08	52.92	49.00	37.13	32.37
Ours+QDROP	2/4	**66.18** (+1.52)	**70.53** (+0.45)	**57.85** (+4.93)	**53.71** (+4.71)	**40.09** (+2.96)	**35.85** (+3.48)
AdaQuant [10]	3/3	60.09	67.46	2.23	-	-	-
AdaRound* [21]	3/3	63.91	64.85	34.55	18.16	8.13	4.45
BRECQ* [16]	3/3	64.83	70.06	52.03	45.54	29.79	25.52
Ours+BRECQ	3/3	**65.64** (+0.81)	**70.68** (+0.62)	**57.14** (+5.11)	**50.21** (+4.67)	**35.11** (+5.32)	**30.26** (+4.74)
QDROP [30]	3/3	65.56	71.07	54.27	49.26	35.14	29.40
Ours+QDROP	3/3	**66.30** (+0.74)	**71.92** (+0.85)	**58.40** (+4.13)	**51.78** (+2.52)	**38.43** (+3.29)	**32.96** (+3.56)
BRECQ* [16]	2/2	46.89	40.18	7.03	5.60	1.87	1.62
Ours+BRECQ	2/2	**52.02** (+5.13)	**43.72** (+3.54)	**13.84** (+6.81)	**9.46** (+3.86)	**3.43** (+1.56)	**3.22** (+1.6)
QDROP [30]	2/2	51.14	54.74	8.46	8.67	3.31	2.77
Ours+QDROP	2/2	**54.46** (+3.32)	**56.82** (+2.08)	**14.44** (+5.98)	**11.40** (+2.73)	**4.18** (+0.87)	**3.09** (+0.32)

Comparison of MRECG with state-of-the-art methods. MRECG shows certain superiority on different models with different bit configurations. In particular, MRECG has huge gains on small models

References

[1] Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. {BRECQ}: Pushing the limit of post-training quantization by block reconstruction. ICLR 2021

[2] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In International Conference on Machine Learning, PMLR, 2020.

Thanks