Symmetry Regularization and Saturating Nonlinearity for Robust Quantization Supplementary Materials

1 Error Propagation Comparison

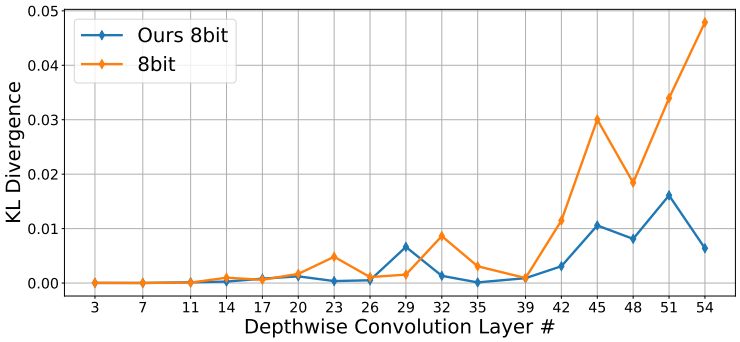


Fig. 1A: KL Divergence of depthwise convolution output between the baseline MobileNet-V2 and the model with ours (e.g., SymReg, SatNL, and ASAM) on the ImageNet dataset. The weights of both networks are quantized into 8-bit through PTQ (ACIQ).

Fig. 1A shows the layer-wise KL divergence of output activation before and after the 8-bit weight quantization. As shown in the figure, the KL divergence of the baseline network becomes larger in the last depthwise convolution layer, while the divergence of the proposed network has a much smaller difference. Because the proposed methods minimize the quantization error through SatNL, the layer-wise error is smaller than the original network. In addition, SymReg mitigates the error propagation, which prevents the accumulation of quantization errors over multiple layers. As a result, the output activation could maintain the consistent features, and we could enjoy the benefit of low-precision computation with minimal accuracy degradation.

2 Weight Distribution Visualization

Fig. 2A shows the effect of the proposed methods by visualizing the histograms of weights in the 15th convolution layer of MobileNet-v2 at CIFAR-100. As shown in the left figure, the original weight has an irregular distribution with a few

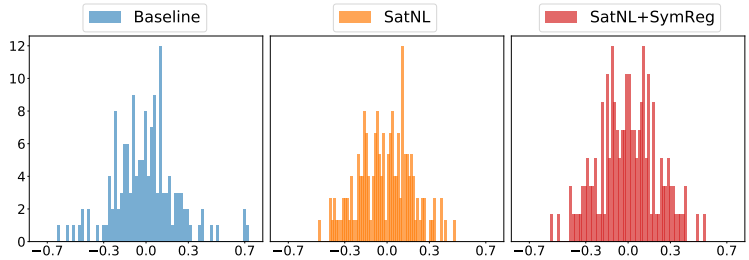


Fig. 2A: Weight distribution of convolution kernels in the 15th convolution layer of MobileNet-V2 at CIFAR-100. Left : Baseline, Middle : SatNL, Right : SatNL+SymReg.

large values. Due to these infrequent values, the quantization error is increased after the quantization. When we train the network with SatNL, the distributions are concentrated within the narrowed range, as shown in the middle figure. As a result, the statistics difference before and after the quantization could be minimized. After applying SymReg in addition to SatNL, the distribution now becomes symmetric, and thereby the biased quantization error is forced to zero regardless of quantization algorithms. While the proposed methods reduce the degree of freedom of weight, introducing minor accuracy degradation, the robustness of the network against quantization could be enhanced significantly.

3 Training Configurations

Table 1A: Hyper-parameters to train the networks

					Cosine annealing with warmup		ASAM
configuration		epoch	lr	weight decay	warmup len	η_{min}	ρ
ResNet18	ImageNet	150	0.4	1×10^{-5}	5	1×10^{-2}	1
MobileNetv2	Cifar100	120	0.4	5×10^{-5}	5	1×10^{-2}	1
	ImageNet	150	0.4	1×10^{-5}	5	1×10^{-2}	1
MobileNetv3	ImageNet	240	0.4	1×10^{-5}	5	1×10^{-2}	0.2

In this work, we need to train the target models from scratch for PTQ experiments. Tab. 1A shows the hyper-parameters we use to train the models. We use the well-known SGD with a momentum algorithm and the exponential moving average of parameters. In the case of QAT, we apply 90 epochs of fine-tuning with 1/10 lower learning rates than used in the pre-training stage. The rest of the hyper-parameters are set identical to the initial full-precision pre-training stage.

4 Reproducibility

The entire source code is available in the author's public Github repository. https://github.com/EunhyeokPark/RobustQuant

5 Other Non-linearities for SatNL

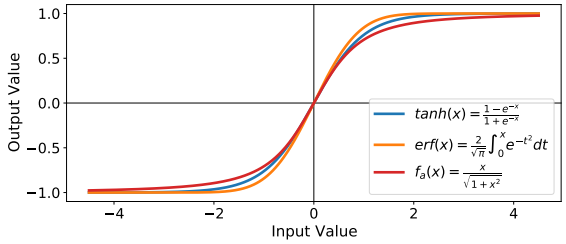


Fig. 3A: Various saturating non-linearity functions.

As mentioned in section 4.2, SatNL requires three properties; 1. odd function, 2. bounded range, and 3. decreased slope. We reproduce the experiments of Table 1 in the main paper by replacing *tanh* with two other functions satisfying the properties (Fig. 3A). We observe that the accuracy difference is negligible, proving that the choice of the SatNL function has no notable impact on the accuracy.

6 ImageNet Experiments with KURE

Tab. 2A extends Table 1 in the main paper by including the comparison to the previous state-of-the-art method, KURE[3]. As shown in the table, the joint regularization of ours and KURE shows the highest accuracy in extreme low-precision PTQ (3-bit) compared to other methods. However, because KURE is a strong regularization, it introduces slight accuracy degradation in the full-precision pre-training stage. When applying PTQ with 4-bit or higher precision, ours without KURE shows higher accuracy in every case. Our novel ideas push the boundaries of achievable accuracy when one quantizes the network with PTQ in 4-bit or higher.

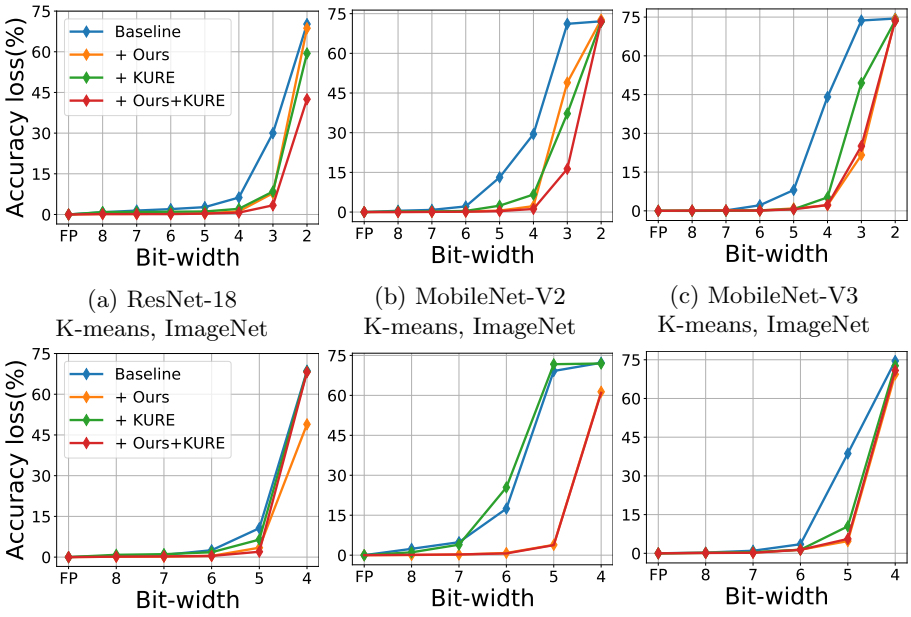
7 Additional Experiments on Non-linear PTQ Algorithms

In order to show the outstanding benefit of the proposed methods for the non-linear quantization algorithms, we conducted extensive studies based on non-linear PTQ algorithms (i.e., logarithm-based quantization and K-means clustering-based quantization). As shown in Fig. 4A, our proposed methods increase the

Table 2A: Results of applying PTQ to baseline and network with proposed ideas including KURE on ImageNet dataset. The values in the table represent the top-1 accuracy. The dashed cells represent the points where the PTQ fails to converge, having lower than 1 % of accuracy.

Model	PTQ	Method	Weight/activation bit-width configuration							
			FP	4/FP	3/FP	2/FP	6/6	5/5	4/4	3/3
ResNet-18	ACIQ	Baseline	70.54	47.44	-	-	68.70	64.87	38.46	-
		Ours	70.92	69.22	49.06	-	70.02	68.99	66.65	42.95
		KURE	69.39	66.69	44.19	-	68.01	66.84	61.80	26.58
		Ours+KURE	70.33	69.85	67.59	-	69.23	68.65	67.11	61.07
	AdaQuant	Baseline	70.54	69.29	66.18	3.23	70.17	69.55	67.67	57.57
		Ours	70.92	70.36	68.84	48.39	70.75	70.37	69.35	64.04
		KURE	69.39	69.09	68.32	62.21	69.23	68.77	67.77	64.22
		Ours+KURE	70.33	69.96	69.21	63.16	70.11	69.77	69.14	66.06
	QDrop	Baseline	70.54	70.15	69.39	66.40	70.27	69.93	68.91	65.75
		Ours	70.92	70.69	70.06	66.98	70.81	70.57	69.93	67.45
		KURE	69.39	69.35	69.14	67.60	69.25	69.13	68.30	66.10
		Ours+KURE	70.33	70.18	69.86	67.77	70.13	69.86	69.38	67.47
MobileNet-V2	ACIQ	Baseline	72.22	28.68	-	-	69.30	64.20	18.15	-
		Ours	72.87	70.07	40.79	-	71.07	68.66	58.30	6.25
		KURE	72.07	54.34	6.43	-	69.77	64.37	39.31	2.14
		Ours+KURE	72.48	70.30	42.24	-	70.68	68.31	61.51	13.87
	AdaQuant	Baseline	72.22	70.67	59.80	-	71.52	70.72	63.70	-
		Ours	72.87	72.23	69.03	-	72.27	71.76	68.91	18.36
		KURE	72.07	71.51	68.71	3.58	71.62	70.71	66.25	4.61
		Ours+KURE	72.48	71.93	70.17	10.16	71.90	71.26	68.84	29.27
	QDrop	Baseline	72.22	71.41	68.32	48.68	71.57	70.64	67.08	50.79
		Ours	72.87	72.44	71.18	61.68	72.61	72.05	69.87	62.55
		KURE	72.07	71.75	70.55	59.51	71.76	70.91	68.23	56.20
		Ours+KURE	72.48	72.13	71.25	63.78	72.15	71.60	70.00	63.30
MobileNet-V3	ACIQ	Baseline	74.52	29.65	-	-	-	-	-	-
		Ours	74.43	61.95	1.04	-	-	-	-	-
		KURE	73.81	55.15	3.09	-	-	-	-	-
		Ours+KURE	73.84	66.21	5.44	-	-	-	-	-
	AdaQuant	Baseline	74.52	72.92	64.17	-	72.73	68.95	43.88	-
		Ours	74.43	73.51	70.50	2.87	72.69	71.02	62.73	-
		KURE	73.81	73.11	70.46	7.44	72.63	70.63	59.25	-
		Ours+KURE	73.84	73.25	70.91	20.05	72.29	70.32	62.35	1.84

robustness by a large margin in both methods, allowing minimal accuracy degradation in low-precision. This result verifies that the proposed methods are also applicable for non-linear quantization. Compared to the previous best, KURE, ours gives comparable or slightly better robustness in the optimized networks, i.e., MobileNet-V2/V3. According to our observation, SatNL is highly beneficial to stabilize the non-linear PTQ process because the statistical difference before and after the quantization could be minimized.



(a) ResNet-18 K-means, ImageNet

(b) MobileNet-V2 K-means, ImageNet

(c) MobileNet-V3 K-means, ImageNet

(d) ResNet-18 Log-Quant, ImageNet

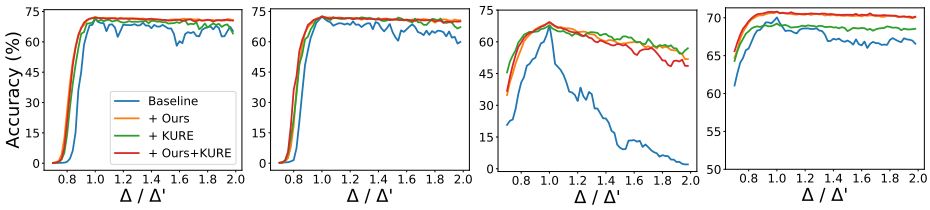
(e) MobileNet-V2 Log-Quant, ImageNet

(f) MobileNet-V3 Log-Quant, ImageNet

Fig. 4A: Robustness of quantized networks with non-linear quantization algorithms. The weight precision is changed while the activation remains full-precision.

8 Addtional Results Regarding Robustness for Quantization Step Size

Fig. 5A shows the additional experiments for measuring the robustness of networks for step size changes corresponding to Fig. 7 in the main paper. In all cases, the quantized models with proposed methods maintain the accuracy in various quantization step sizes. Our methods are beneficial for robust quantization even for the optimized networks, i.e., MobileNet-V2/V3.



(a) MobileNet-V2 PTQ 6b, ImageNet (b) MobileNet-V3 PTQ 6b, ImageNet (c) ResNet-18 PTQ 4b, ImageNet (d) ResNet-18 PTQ 6b, ImageNet

Fig. 5A: Robustness of quantized network when we change step size of quantization operator for weight. The networks are optimized for the step size Δ', and the accuracy is measured with the scaled step size Δ. All networks are quantized into the given bitwidth with PTQ [2], including activation and weight.

9 Explanation of Equations

9.1 Equation 4 and 6

In Equations 4 and 6, we follow the derivation of ACIQ [1] regarding the expected mean-square-error of linear quantization. When we apply b-bit quantization to the quantization boundaries $[-\alpha, \alpha]$, the quantization interval is equally divided into 2^b discrete levels. When the density function is given as $f(x)$, the overall quantization error is expressed as follows:

$$\begin{aligned}
&\text{Quantization Error} = E[(W - Q(W))^2] \\
&= \overbrace{\int_{-\infty}^{\infty} f(x) \cdot (x-\alpha)^2 dx}^{\text{quantization error}} \\
&= \overbrace{\cdot \int_{\alpha}^{\infty} f(x) \cdot (x-\alpha)^2 dx + \cdot \int_{-\infty}^{\alpha} f(x) \cdot (x-\alpha)^2 dx}^{\text{truncation error}} \\
&+ \overbrace{\sum_{i=0}^{2^M-1} \int_{-\alpha+i\cdot\Delta}^{-\alpha+(i+1)\cdot\Delta} f(x) \cdot (x-q_i)^2 dx}^{\text{rounding error}},
\end{aligned} \tag{1}$$

where $\Delta = 2 \cdot \alpha / 2^b$ and $q_i = -\alpha + (2i+1) \cdot \Delta/2$.

In the previous study [1], the rounding error is approximated by a piece-wise linear function based on the slope and the value of the density function at the midpoint of quantization levels, q_i. The rounding error of the quantization noise is approximated as follows:

$$\overbrace{\sum_{i=0}^{2^M-1} \int_{-\alpha+i\cdot\Delta}^{-\alpha+(i+1)\cdot\Delta} f(x) \cdot (x-q_i)^2 dx}^{\text{rounding error}} \approx \frac{\alpha^2}{3 \cdot 2^{2b}}. \tag{2}$$

By substituting the above equation to the rounding error term, the quantization error is summarized as follows:

$$\begin{aligned}
&\text{Quantization Error} = E[(W - Q(W))^2] \\
&\approx \overbrace{\cdot \int_{\alpha}^{\infty} f(x) \cdot (x-\alpha)^2 dx + \cdot \int_{-\infty}^{-\alpha} f(x) \cdot (x-\alpha)^2 dx}^{\text{truncation error}} + \overbrace{\frac{\alpha^2}{3 \cdot 2^{2b}}}^{\text{rounding error}},
\end{aligned} \tag{3}$$

where α is the truncation boundary that minimizes $||W - Q(W)||_2$. In addition, because Gaussian distribution is an even function, two terms of truncation error are identical. Overall, the quantization error is summarized as follows:

$$\begin{aligned}
&\text{Quantization Error} = E[(W - Q(W))^2] \\
&\approx \overbrace{2 \cdot \int_{\alpha}^{\infty} f(x) \cdot (x-\alpha)^2 dx}^{\text{truncation error}} + \overbrace{\frac{\alpha^2}{3 \cdot 2^{2b}}}^{\text{rounding error}}.
\end{aligned} \tag{4}$$

In the case of Equation 6, the probability density function of $G(x)$ is the same as $f(x)$, therefore the quantization error in $[-d, d]$ could be achievable following

the similar derivation of Equation 10 as given by:

$$\begin{aligned}&\text{Quantization Error'} \in (-d, d)\\ &\approx \overbrace{2 \cdot \int_{\alpha'}^{d} f(x) \cdot (x-\alpha')^2 dx}^{\text{truncation error}} + \overbrace{\frac{\alpha'^2}{3 \cdot 2^{2b}}}^{\text{rounding error}},\end{aligned} \tag{5}$$

where α' is the truncation boundary that minimizes $||W' - Q(W')||_2$. In addition, the error of clamped values is expressed as:

$$\begin{aligned}&F(-|d|)(d-\alpha')^2 + (1-F(d))(d-\alpha')^2\\ &= 2 \cdot F(-|d|) \cdot (d-\alpha')^2\end{aligned} \tag{6}$$

By combining the above two terms, we can get the overall quantization errors of the clamped weight:

$$\begin{aligned}&\text{Quantization Error}' = E[(W' - Q(W'))^2]\\ &\approx \overbrace{2 \cdot \Big(F(-|d|) \cdot (d-\alpha')^2 + \int_{\alpha'}^{d} f(x) \cdot (x-\alpha')^2 dx\Big)}^{\text{truncation error}} + \overbrace{\frac{\alpha'^2}{3 \cdot 2^{2b}}}^{\text{rounding error}}.\end{aligned} \tag{7}$$

9.2 Proof of Error Comparison

When we compare Eq. (4) and Eq. (6), Eq. (6) always has a smaller error than Eq. (4), showing that the clamped distribution is more robust than the unbounded distribution. In order to prove the relationship mentioned above, we will first give two lemmas.

Lemma 1. For the arbitrary $A < d$, the quantization error of unbounded distribution with truncation boundary A is always larger than the quantization error of bounded distribution with truncation boundary A.

The difference of error is given as:

$$\begin{aligned}&\Delta\text{Error}\\ &= 2 \cdot \int_{A}^{\infty} f(x) \cdot (x-A)^2 dx - 2 \cdot F(-|d|) \cdot (d-A)^2\\ &= 2 \cdot \Big(\int_{A}^{\infty} f(x) \cdot (x-A)^2 dx - \int_{d}^{\infty} f(x) \cdot (d-A)^2 dx\Big)\\ &= 2 \cdot \int_{A}^{d} f(x) \cdot (x-A)^2 dx\\ &+ 2 \cdot \int_{d}^{\infty} f(x)\big((x-A)^2 - (d-A)^2\big) dx.\end{aligned} \tag{8}$$

The last two terms are always positive, therefore lemma 1 holds.

Lemma 2. For the arbitrary $A < d$, the quantization error of bounded distribution with truncation boundary A is always larger than or equal to the quantization error of bounded distribution with truncation boundary α'.

From the definition of α', where α' is the truncation boundary that minimizes $||W' - Q(W')||_2$, lemma 2 is always valid.

From lemma 1 and lemma 2, for the arbitrary $A < d$, the quantization error of unbounded distribution with truncation boundary A is larger than the quantization error of bounded distribution with truncation boundary α'. Therefore, Eq.(6) is always smaller than to Eq.(4). ■

References

1. Banner, R., Nahshan, Y., Hoffer, E., Soudry, D.: ACIQ: analytical clipping for integer quantization of neural networks. CoRR **abs/1810.05723** (2018), http://arxiv.org/abs/1810.05723
2. Hubara, I., Nahshan, Y., Hanani, Y., Banner, R., Soudry, D.: Accurate post training quantization with small calibration sets. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 4466–4475. PMLR (2021), http://proceedings.mlr.press/v139/hubara21a.html
3. Shkolnik, M., Chmiel, B., Banner, R., Shomron, G., Nahshan, Y., Bronstein, A.M., Weiser, U.C.: Robust quantization: One model to rule them all. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H.T. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/3948ead63a9f2944218de038d8934305-Abstract.html