Improving Low-Precision Network Quantization via Bin Regularization Supplementary Material

Tiantian Han Dong Li Ji Liu Lu Tian Yi Shan

Xilinx Inc., Beijing, China

{hantian, dongl, jiliu1, lutian, yishan}@xilinx.com

1. Overview

In this supplementary material, we present more detailed experimental results and analysis.

- We conduct ablation study on our two-stage optimization strategy.
- We evaluate the impact of loss weight λ for different bit widths.
- We evaluate variability of results for different bit widths on MobileNetV3-Small.
- We analyze the quantization error of our method for different types of low-bit networks.
- We show more examples of data distributions with the LSQ baseline, KURE regularization and our BR method.

2. Optimization Strategy

We adopt a two-stage optimization strategy by first updating the step size with m epochs and then adding the bin regularization term. We train each low-precision network for the same 90 epochs in total. Table 1 presents ablation experiments for MobileNetV2 with different bit widths. The results show that jointly updating the step size and regularizing quantization bins from the beginning of training (i.e., $m = 0$) performs worse than the two-stage strategy for all bit widths. Besides, we find that adding bin regularization too late may produce sub-optimal performance for 2-bit MobileNetV2, e.g., 47.6% ($m = 60$) vs. 50.6% ($m = 30$).

3. Parameter Analysis

We analyze the impact of loss weight with different bit widths on MobileNetV3-Small. Table 2 shows that different choices of λ in a reasonable range works well (e.g., $\lambda = 2, 1, 0.5$). Too small loss weight of bin regularization may yield decreased accuracy (e.g., 32.7% top-1 accuracy with $\lambda = 0.05$ for 2-bit network).

m epochs	W/A	Acc@1 (%)	Acc@5 (%)
Full-precision	32/32	71.8	90.2
LSQ* [1]	4/4	69.5	89.2
0	4/4	69.5	89.0
30	4/4	**70.4**	89.4
45	4/4	69.6	89.1
60	4/4	70.3	**89.5**
LSQ* [1]	3/3	65.3	86.3
0	3/3	66.0	86.7
30	3/3	67.4	87.4
45	3/3	67.3	87.4
60	3/3	**67.5**	**87.7**
LSQ* [1]	2/2	46.7	71.4
0	2/2	44.3	69.1
30	2/2	**50.6**	**74.6**
45	2/2	47.8	72.2
60	2/2	47.6	72.0

Table 1: Ablation study on our two-stage optimization strategy for MobileNetV2 with different bit widths. “*” means our re-implementation.

4. Variability of the results

To evaluate the variability of results, we conduct 5 trials for different low-bit models on MobileNetV3-Small. Table 3 shows that the variation of our experimental results is small.

5. Analysis of Quantization Error

We calculate the MSE quantization error (MSE-QE) and Mean Bin Loss (MBL) for different low-bit networks in Table 4. Taking 2-bit as example, our method can achieve the smallest MSE quantization error and best accuracy compared to the LSQ baseline and KURE regularization method. Although our method does not explicitly optimize

λ	W/A	Acc@1 (%)	Acc@5 (%)
Full-precision	32/32	65.1	85.4
10	4/4	61.0	82.5
5	4/4	61.4	82.6
2	4/4	61.5	**82.8**
1	4/4	**61.7**	**82.8**
0.5	4/4	61.6	83.1
4	3/3	55.2	78.1
2	3/3	55.5	78.4
1	3/3	55.8	78.7
0.5	3/3	**56.0**	**78.8**
0.25	3/3	55.8	**78.8**
0.05	3/3	54.7	77.8
8	2/2	35.4	60.0
4	2/2	**36.3**	60.6
2	2/2	36.1	60.9
1	2/2	35.6	60.3
0.5	2/2	**36.3**	**61.0**
0.25	2/2	35.0	59.5
0.05	2/2	32.7	56.6

Table 2: Impact of loss weight λ on MobileNetV3-Small with different bit widths.

Method	W/A	Acc@1(%)	Acc@5(%)
Full-precision	32/32	65.1	85.4
LSQ*	4/4	61.0	82.6
LSQ + BR (Ours)	4/4	61.5±0.1	82.8±0.2
LSQ*	3/3	52.0	76.1
LSQ + BR (Ours)	3/3	56.0±0.1	78.8±0.3
LSQ*	2/2	31.4	55.5
LSQ + BR (Ours)	2/2	36.3±0.3	61.0±0.5

Table 3: Variability of the results on MobileNetV3-Small. "*" means our re-implementation.

Network	Methods	W/A	MSE-QE	MBL	Acc@1 (%)
ResNet18	LSQ* [1]	2/2	3.0e-04	1.2e-03	66.5
	KURE [2]	2/2	**2.0e-04**	7.0e-04	66.2
	BR (Ours)	2/2	**2.0e-04**	**6.0e-04**	**67.2**
MobileNetV2	LSQ* [1]	2/2	6.1e+02	1.0e+04	46.7
	KURE [2]	2/2	5.7e+15	7.9e+16	37.0
	BR (Ours)	2/2	**2.5e-02**	**1.6e-01**	**50.6**
MobileNetV3-Small	LSQ* [1]	2/2	7.4e-03	2.9e-03	31.4
	BR (Ours)	2/2	**5.0e-04**	**1.7e-03**	**36.3**

Table 4: Analysis of quantization error on different types of low-bit networks.

the overall quantization error, we find that the error still drops after regularizing each bin.

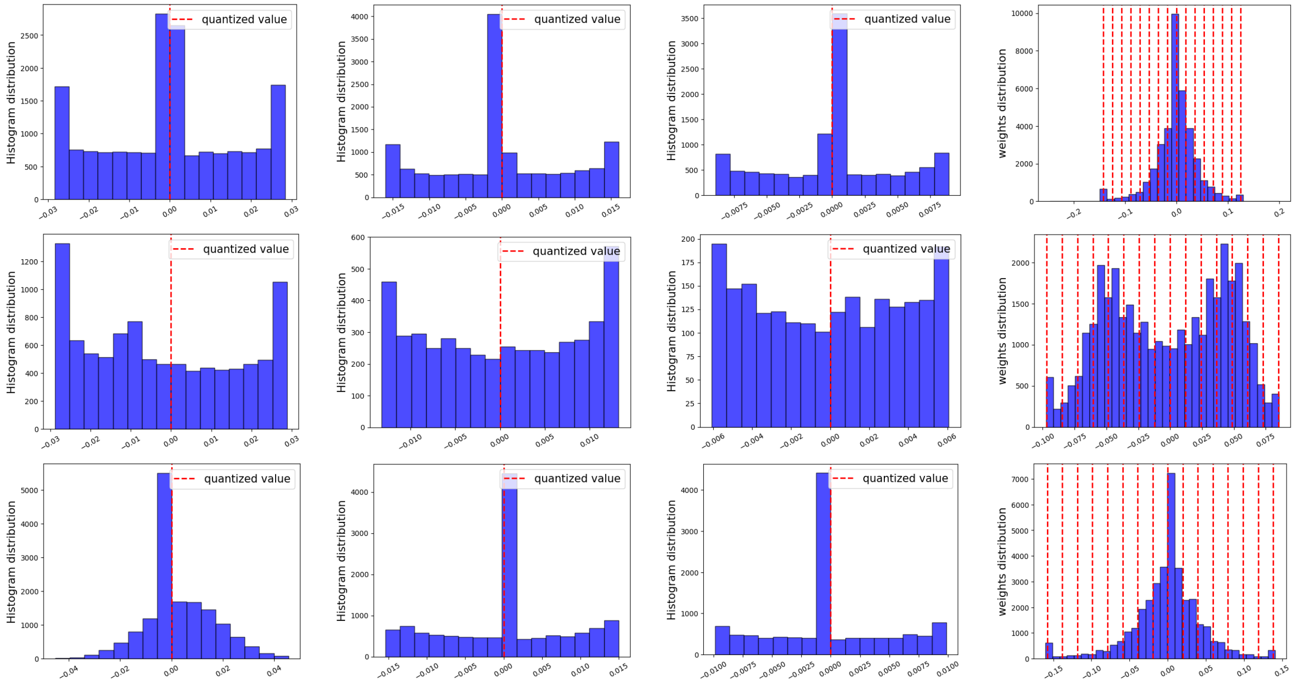
6. Visualization of Bin Distribution

Figure 1 and 2 show bin and global distributions for different bit widths on ResNet18 and MobileNetV3-Small, respectively. Figure 3, 4, 5 show bin and global distributions for different network layers for ResNet18, MobileNetV2, MobileNetV3-Small, respectively. The LSQ baseline neither explicitly constrain the overall nor the bin distributions. KURE regularization encourages the overall distribution to be uniform, while our method encourages the bin distribution to be sharp. The curves show that the quantized values in each bin are concentrated around the target value (red dash line in the figures) by our BR method, which validate our idea.

References

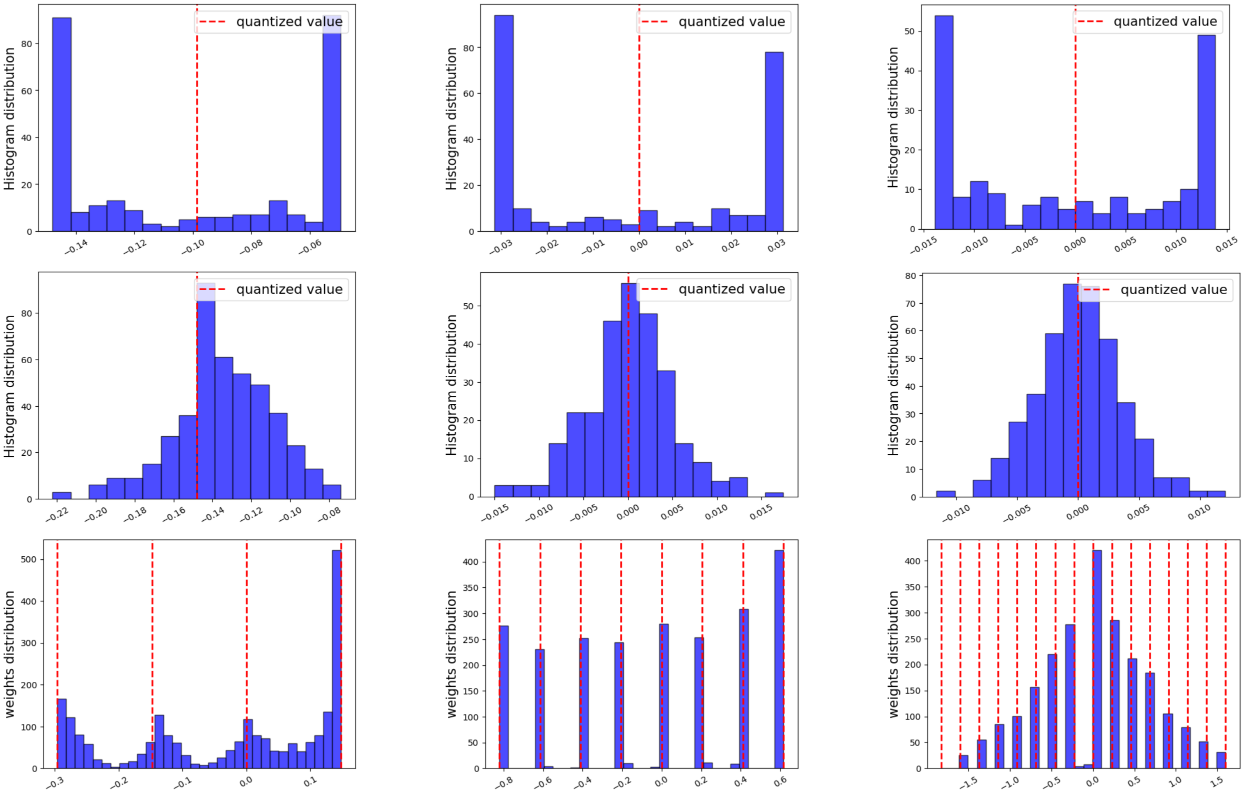
[1] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *ICLR*, 2020. 1, 2

[2] Moran Shkolnik, Brian Chmiel, Ron Banner, Gil Shomron, Yuri Nahshan, Alex Bronstein, and Uri Weiser. Robust quantization: One model to rule them all. In *NeurIPS*, 2020. 2



(a) 2-bit (bin) (b) 3-bit (bin) (c) 4-bit (bin) (d) 4-bit (global)

Figure 1: Bin distributions with different bit widths (a, b, c) and 4-bit global distribution (d) on `layer2` of ResNet18. Row 1∼3 represent the original LSQ, LSQ+KURE, and LSQ+BR (Ours), respectively.

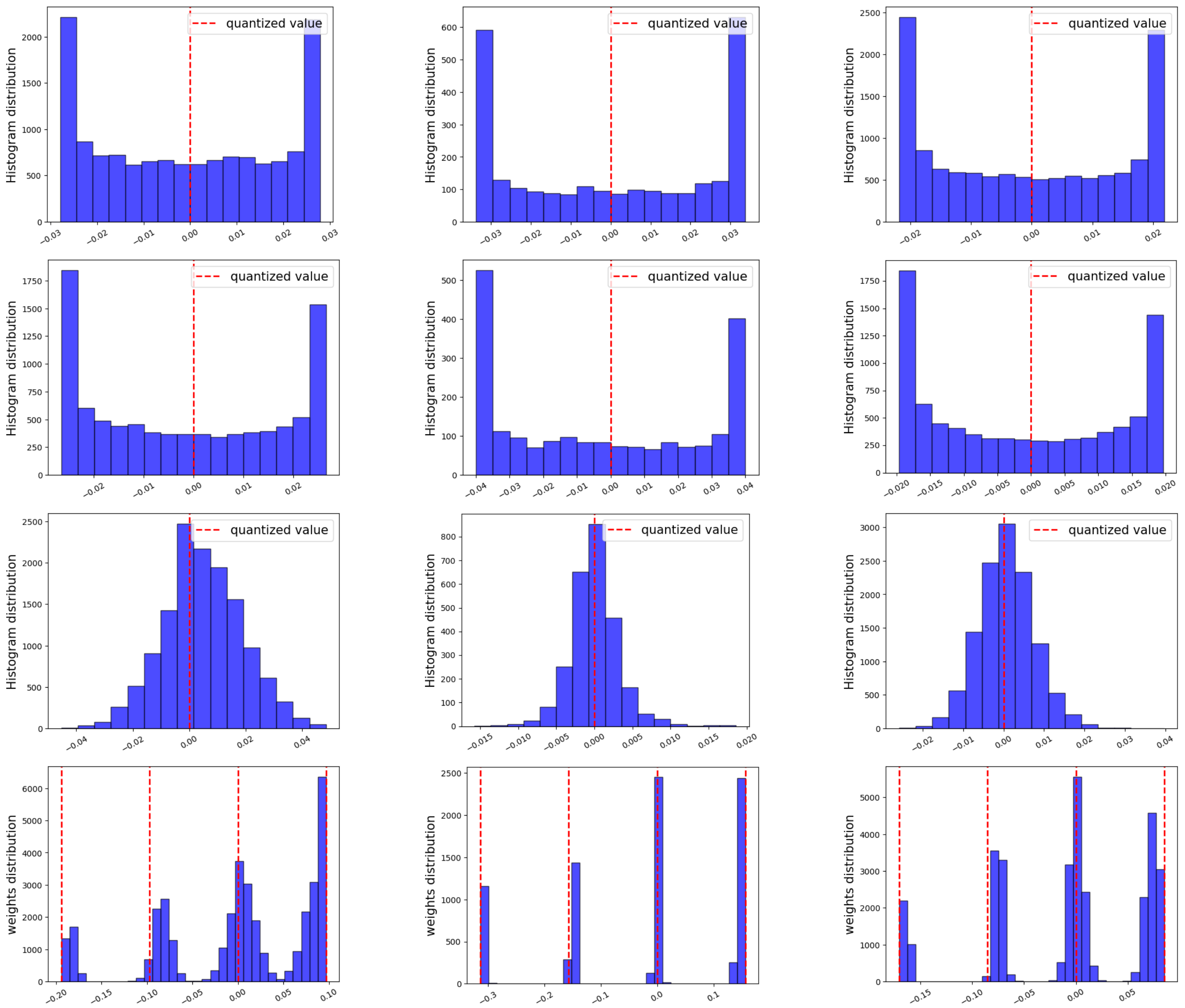


(a) 2-bit

(b) 3-bit

(c) 4-bit

Figure 2: Bin and global distributions with different bit widths on `layer12` of MobileNetV3-Small. The first two rows represent the bin distributions of original LSQ and LSQ+BR (Ours), respectively. The last row represents the global distribution of LSQ+BR (Ours).

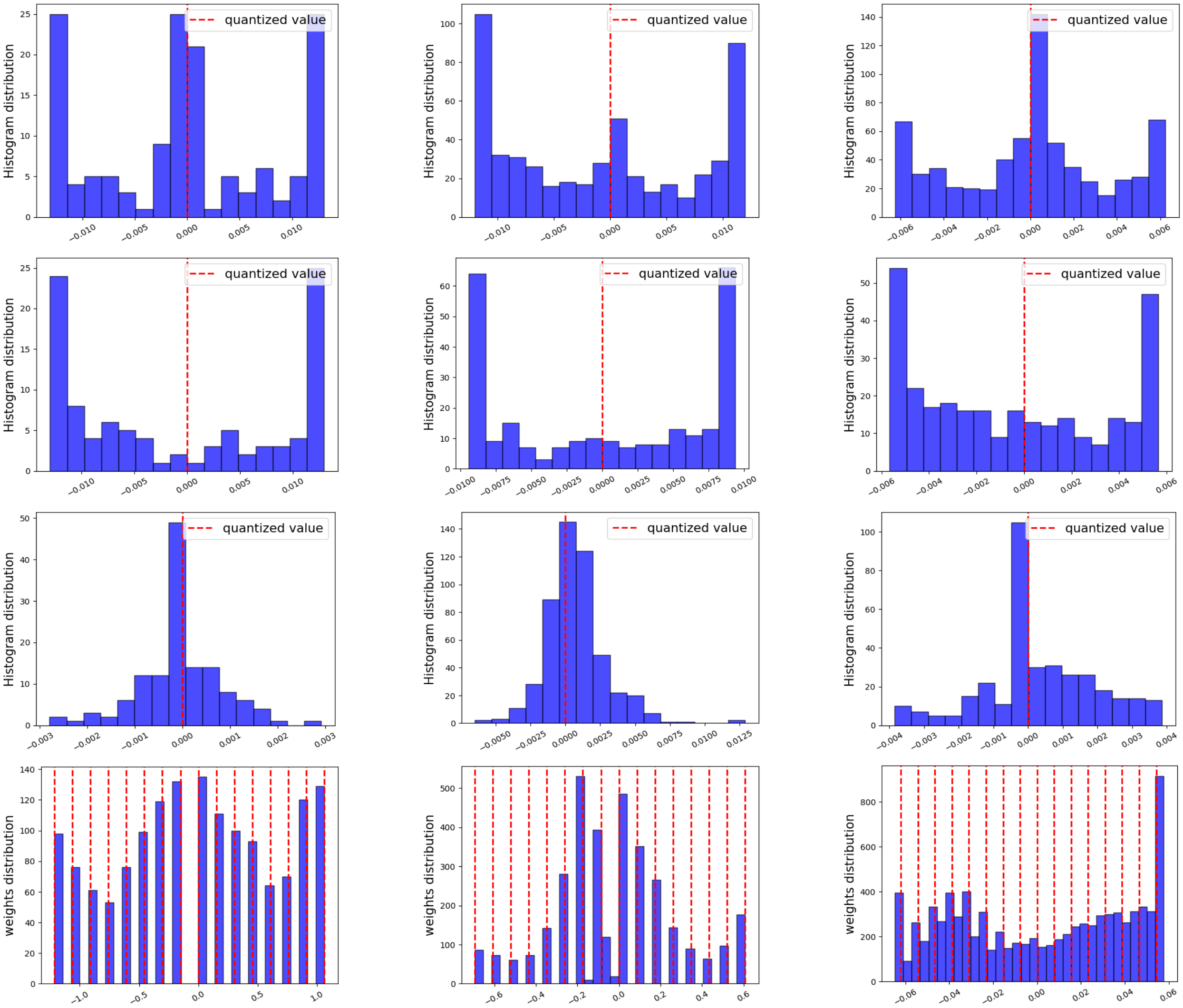


(a) `layer3`

(b) `layer7`

(c) `layer12`

Figure 3: Bin and global distributions for different network layers of 2-bit ResNet18. The first three rows represent the bin distributions of the original LSQ, LSQ+KURE and LSQ+BR (Ours), respectively. The last row represents the global distribution of LSQ+BR (Ours).

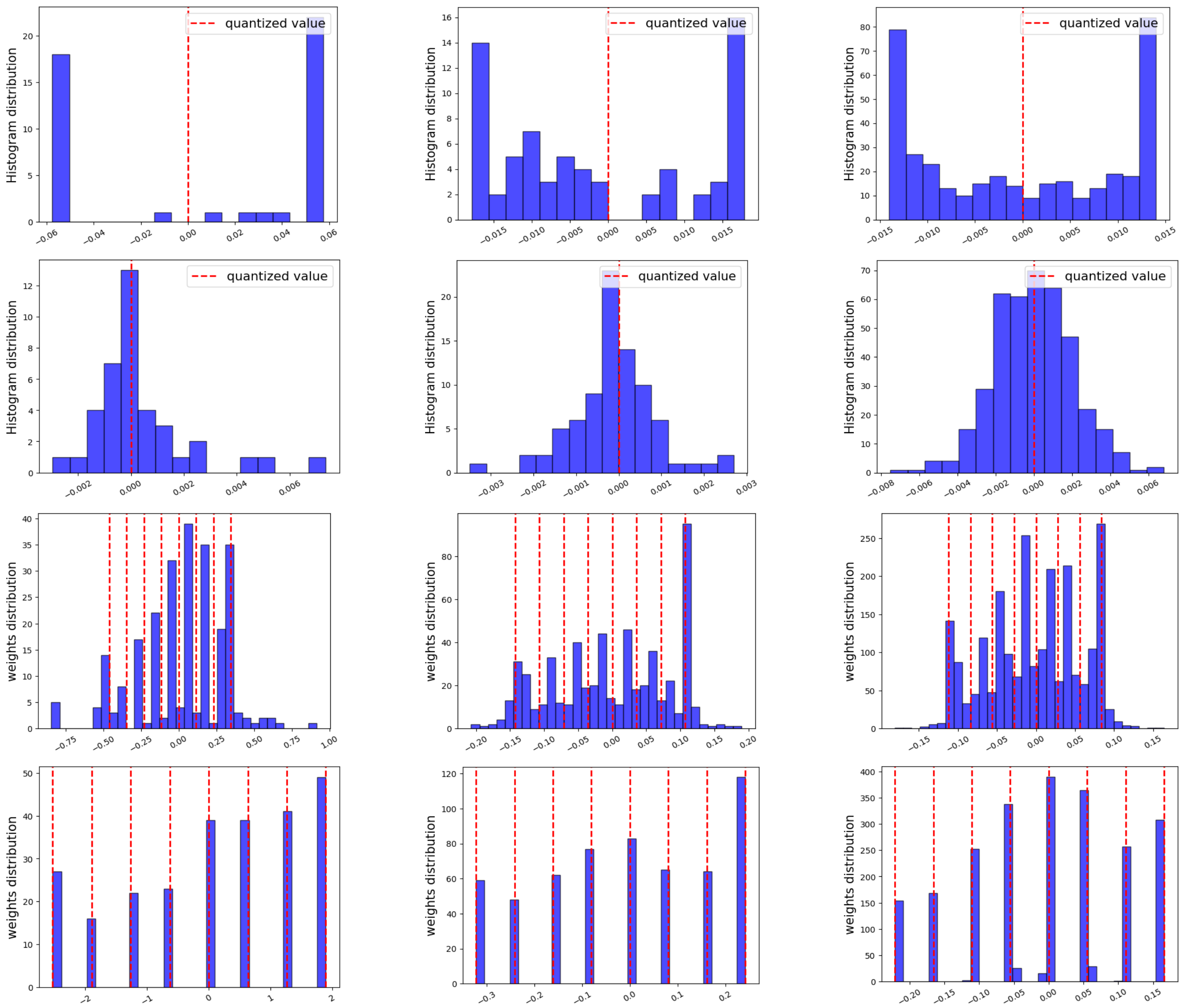


(a) `layer3`

(b) `layer22`

(c) `layer46`

Figure 4: Bin and global distributions for different network layers of 4-bit MobileNetV2. The first three rows represent the bin distributions of the original LSQ, LSQ+KURE and LSQ+BR (Ours), respectively. The last row represents the global distribution of LSQ+BR (Ours).



(a) `layer2`

(b) `layer23`

(c) `layer40`

Figure 5: Bin and global distributions for different network layers of 3-bit MobileNetV3-Small. The first two rows represent the bin distributions of the original LSQ and LSQ+BR (Ours), respectively. The last two rows represent the global distributions of the original LSQ and LSQ+BR (Ours), respectively.