Diversifying Sample Generation for Accurate Data-Free Quantization Supplemental Material

Xiangguo Zhang*[1], Haotong Qin*[1], Yifu Ding[1], Ruihao Gong[3,4],
Qinghua Yan[1], Renshuai Tao[1], Yuhang Li[2], Fengwei Yu[3,4], Xianglong Liu[1†]
[1]Beihang University [2]Yale University [3]SenseTime Research [4]Shanghai AI Laboratory
{xiangguozhang, zjdyf, yanqh, rstao}@buaa.edu.cn, yuhang.li@yale.edu,
{qinhaotong, xlliu}@nlsde.buaa.edu.cn, {gongruihao, yufengwei}@sensetime.com

1. Visualizations of Activation Statistics from Real and Synthetic Data

In our paper, we have presented the distributions of mean and standard deviation values in one channel (Figure 3 in our main paper). To give ampler evidence, we provide more visualizations for other channels in Figure 1. Take Figure 1(a) and 1(b) for example, each of them represents the mean or standard deviation in one channel for different types of data. Compared with the statistics of real data which are considered as reasonable references, the mean values of DSG data are more dispersed than those of ZeroQ data, as well as the standard deviation values. Meanwhile, the distributions of ZeroQ in each bar graph are always in the immediate vicinity of BN statistics (the red dash line). Figure 1(c) and 1(d) show the case of another channel, which has just the same phenomenon as described above.

To further demonstrate that to what extent our method really affects, we additionally showcase some box figures in the following for auxiliary instruction (see Figure 2). As can be obviously seen from the figure, both the mean and standard deviation statistics of ZeroQ [1] data are centralized, which have shorter boxes in Figure 2(b) and 2(e). However, real data (Figure 2(a) and 2(d)) and DSG data (Figure 2(c) and 2(f)) have longer boxes, which implies that the distributions of each sample are more dispersed.

Take a closer look at those turning points on the line. Each turning point represents the value of BN mean/standard deviation in one layer. The offset between the turning point and the middle of the box at the same column is much smaller in ZeroQ cases, which means the statistics of ZeroQ data mostly fit the distribution of BN statistics. Whereas as for DSG data, the offset is big, beneath or over the middle of the corresponding box, which implies that the statistics of synthetic data are no longer overfitting to BN statistics.

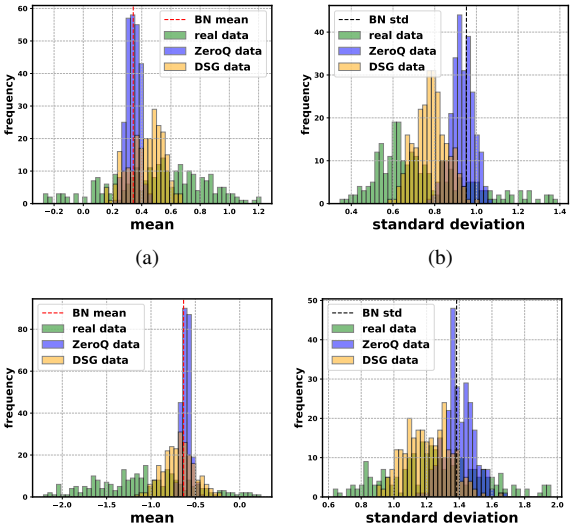

Figure 1: Mean and standard deviation of the activations in two different channels of ResNet-18 when feeding different types of data (with 256 samples). (a) and (b) show the mean and deviation of one specific channel, while (c) and (d) show another.

In short, compared with ZeroQ, distribution statistics of DSG data are much closer to those of real data for two reasons: the dispersion in one layer and the offset to BN statistics. We attribute that to the approaches proposed in our paper, *i.e.* SDA and LSE. The former slacks the alignment of the feature statistics to overcome the overfitting issue and the latter applies the layerwise enhancement to reinforce specific layers. We combine these two approaches and obtain diversified samples.

2. Additional Experiments on Other Dataset and Quantization Methods

In our paper, we have conducted a bunch of experiments to evaluate the effect of our method in both mitigating the

*Equal contribution.
†Corresponding author

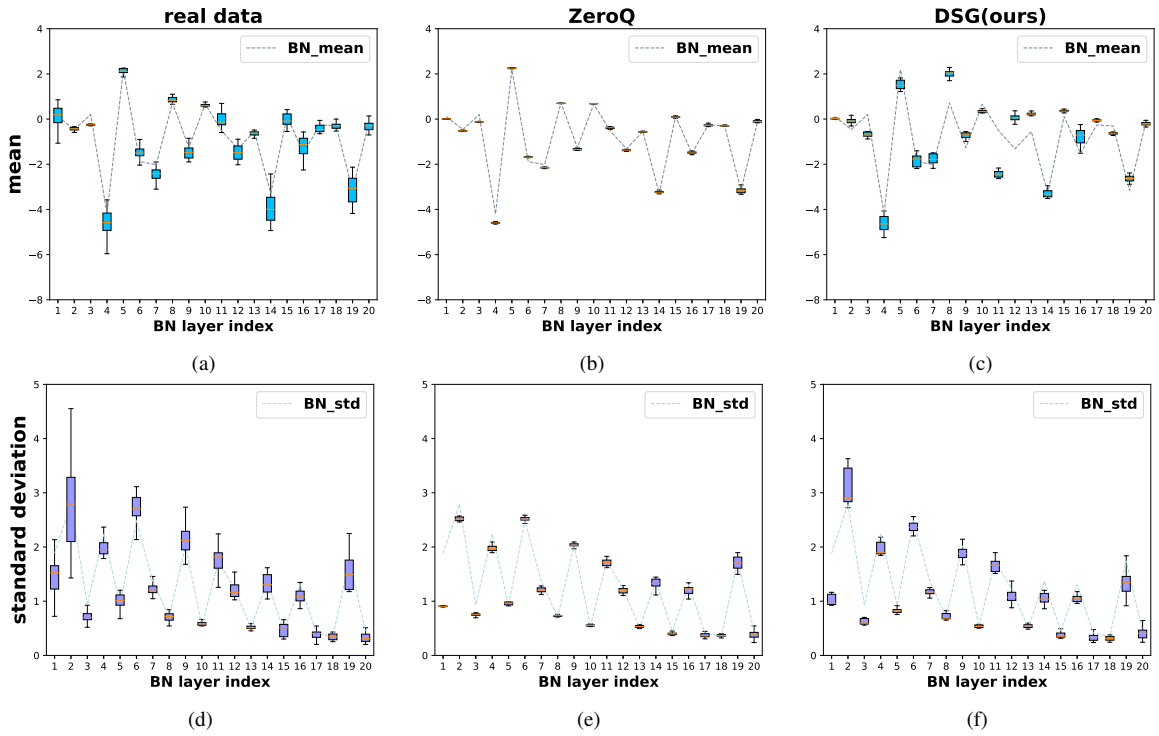

Figure 2: Comparison between real data and synthetic data (generated by DSG and ZeroQ) with 256 samples of each.

homogenization issue and improving the final performance. To further demonstrate the robustness and the general applicability of our method, we provide additional experiments as corroborations to support our viewpoint.

Results on CIFAR-10 We show extra results of our DSG on CIFAR-10 [2] dataset with ResNet-20 [4] and VGG16-bn [8]. See Table 1 and 2. Note that the size of the image sample in the CIFAR-10 dataset is 32×32, much smaller than that in the ImageNet dataset (224×224) which is widely evaluated in our paper. The experimental results show that our DSG still outperforms other SOTA generative data-free quantization methods when generating samples with a small size.

Evaluation with DFQ DFQ [6] has proposed cross-layer range equalization to equalize the different channel ranges of weight in per-layer quantization and bias correction to eliminate the biased quantization error. Both of the two techniques rely on the statistics of BN layers following the convolution layer. Therefore, DFQ only works on specific network architectures and cannot be commonly practiced,

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	94.08
Real Data	✗	✓	4	4	87.38
ZeroQ	✓	✓	4	4	85.39
DSG (Ours)	✓	✓	4	4	**87.75**
Real Data	✗	✓	6	6	93.80
ZeroQ	✓	✓	6	6	93.33
DSG (Ours)	✓	✓	6	6	**93.79**
Real Data	✗	✓	8	8	93.95
ZeroQ	✓	✓	8	8	93.94
DSG (Ours)	✓	✓	8	8	**94.07**

Table 1: Results of ResNet-20 on CIFAR-10.

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	93.86
Real Data	✗	✓	4	4	92.50
ZeroQ	✓	✓	4	4	91.79
DSG (Ours)	✓	✓	4	4	**92.89**
Real Data	✗	✓	6	6	93.48
ZeroQ	✓	✓	6	6	93.45
DSG (Ours)	✓	✓	6	6	**93.68**
Real Data	✗	✓	8	8	93.59
ZeroQ	✓	✓	8	8	93.53
DSG (Ours)	✓	✓	8	8	**93.61**

Table 2: Results of VGG16-bn on CIFAR-10.

since BN layers in DFQ are always needed to proceed behind each convolution layer to quantize the corresponding activations. Fortunately, generative methods, such as ZeroQ and our DSG, can generate synthetic data for arbitrary architectures, and the statistics of activations can be applied to DFQ replacing the BN statistics. Table 3 shows the closeups of two generative data-free quantization method, *i.e.*, ZeroQ and DSG, based on DFQ. Results show that our DSG outperforms ZeroQ by 0.57% and 3.13% in W6A6 and W8A8 cases.

Method	No D	No FT	W-bit	A-bit	Top-1
Baseline	–	–	32	32	69.76
Real Data	✗	✓	6	6	59.16
ZeroQ	✓	✓	6	6	58.12
DSG (Ours)	✓	✓	6	6	**58.69**
Real Data	✗	✓	8	8	69.22
ZeroQ	✓	✓	8	8	65.75
DSG (Ours)	✓	✓	8	8	**68.88**

Table 3: Evaluation with DFQ using ResNet-18 on ImageNet. We use cross-layer equalization and bias correction proposed by DFQ to perform per-layer quantization.

More Evaluation with AdaRound We present more empirical results on AdaRound [5], and the experiments can be broadly divided into two categories: quantizing the weight to extremely low bit-width and quantizing both weight and activation. We use image prior [9] and labels [3] in these experiments. First, we quantize the weight to 3/4 bit-width based on the practical lightweight MobileNetV2 [7]. As Table 4 shown, DSG surpasses the SOTA generative methods by 34.33% with the weight quantized to 3 bit-width. Meanwhile, we provide results of DSG with AdaRound on ResNet-18 [4] quantized to W4A8 in Table 5, and it also shows that our DSG surpasses ZeroQ by a large margin.

Method	No D	Label	Image Prior	W-bit	A-bit	Top-1
Real Data	✗	✗	✗	3	32	58.13
ZeroQ	✓	✓	✓	3	32	11.07
DSG (Ours)	✓	✓	✓	3	32	**45.40**
Real Data	✗	✗	✗	4	32	68.37
ZeroQ	✓	✓	✓	4	32	56.16
DSG (Ours)	✓	✓	✓	4	32	**58.13**

Table 4: Evaluation with AdaRound using MobileNetV2 on ImageNet.

Method	No D	Label	Image Prior	W-bit	A-bit	Top-1
Real Data	✗	✗	✗	4	8	68.24
ZeroQ	✓	✓	✓	4	8	56.34
DSG (Ours)	✓	✓	✓	4	8	**62.40**

Table 5: Evaluation with AdaRound using ResNet-18 on ImageNet.

References

[1] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13169–13178, 2020. 1

[2] Kostyantyn Filonenko, Robert Wisnovsky, Mohamed Chériet (ecole De, Denis J. Dean, Charles W. Anderson, Yann Lecun, and Corinna Costes. techreport learning multiple layers of features from tiny images, by alex. 2

[3] M. Haroush, I. Hubara, E. Hoffer, and D. Soudry. The knowledge within: Methods for data-free model compression. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8491–8499, 2020. 3

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3

[5] Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization, 2020. 3

[6] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *IEEE ICCV*, 2019. 2

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019. 3

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 2

[9] Hongxu Yin, Pavlo Molchanov, Zhizhong Li, Jose M. Alvarez, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion, 2020. 3