Fine-grained Data Distribution Alignment for Post-Training Quantization (Supplementary Material)

Yunshan Zhong[1,2], Mingbao Lin[3], Mengzhao Chen[2], Ke Li[3], Yunhang Shen[3], Fei Chao[2], Yongjian Wu[3], Rongrong Ji[1,2⋆]

[1]Institute of Artificial Intelligence, Xiamen University.
[2]Media Analytics and Computing Lab, Department of Artificial Intelligence, School of Informatics, Xiamen University. [3]Tencent Youtu Lab.
zhongyunshan@stu.xmu.edu.cn, linmb001@outlook.com, cmzxmu@stu.xmu.edu.cn, {tristanli.sh, shenyunhang01}@gmail.com, fchao@xmu.edu.cn, littlekenwu@tencent.com, rrji@xmu.edu.cn

1 More Visualization

1.1 Visualization of MobileNetV1

The visualization of BNS in different layers of pre-trained MobileNetV1 is shown in Fig. S1.

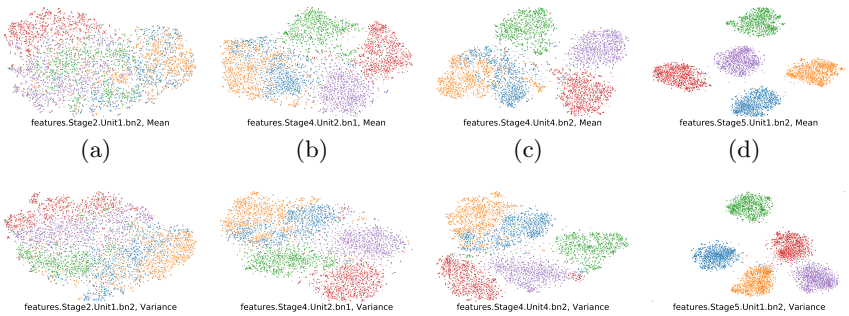

Fig. S1. t-SNE visualization (five classes) of BNS in different layers of pre-trained MobileNetV1 on ImageNet. Best viewed in color.

⋆ Corresponding Author

1.2 Visualization of MobileNetV2

The visualization of BNS in different layers of pre-trained MobileNetV2 is shown in Fig. S2.

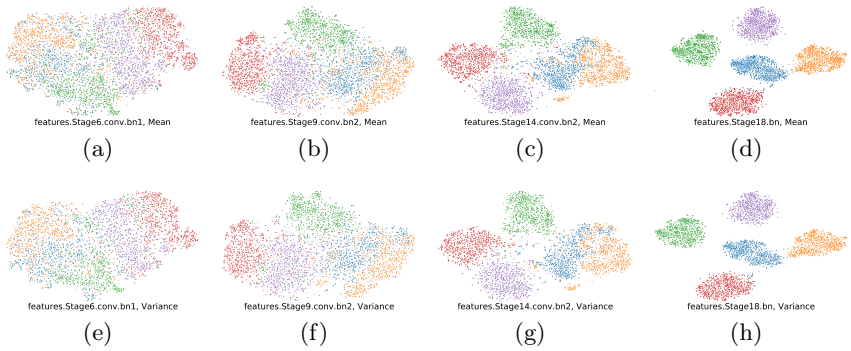

Fig. S2. t-SNE visualization (five classes) of BNS in different layers of pre-trained MobileNetV2 on ImageNet. Best viewed in color.

2 Time cost

Tab. S1 shows that our FDDA and zero-shot quantization (ZSQ) methods have similar training costs. However, our FDDA performs much better as shown in the paper.

Table S1. Training costs of ZSQ and FDDA for 4-bit ResNet-18.

DI/ADI	ZeroQ/DSG	GDFQ	Qimera/ZAQ	FDDA
9.9 hour	6.7 hour	6.9 hour	8.5 hour	7 hour

Table S2. Training costs of PTQ and FDDA for 4-bit ResNet-18.

AdaQuant	LAPQ	Bit-Split	BRECQ	FDDA
0.1 hour	1 hour	3 hour	0.9 hour	7 hour

Tab. S2 reports more training costs from our FDDA than post-training quantization methods. Nevertheless, our FDDA results in a significant performance increase, especially when quantizing small networks such as MobileNetV1 and

Table S3. Accuracy over training time for 4-bit ResNet-18.

Training time (hour)	1	3	5	7
Acc(%)	68.18	68.55	68.74	68.88

MobileNetV2 in the paper. Further, we decrease the training costs (smaller training epochs) and Tab. S3 shows the results. Our FDDA (68.18%) still outperforms the SOTA BRECQ (67.94%) under similar training costs (1 hour *vs.* 0.9 hour). Thus, our FDDA leads to the best performance under the same training cost.

3 Model size and speed

The model size is only related to the specified bits. For example, full-precision ResNet-18 and MobileNetV2 are 11.69MB and 3.5MB, While 4-bit ResNet-18 and MobileNetV2 are 1.47MB and 0.44MB.

After obtaining the quantized model, one can deploy it on hardware with different frameworks depending on the type of hardware. Compared with the full-precision model, the 4-bit model could achieve $\sim$4$\times$ to $\sim$8$\times$ speedups in practice. For example, the 4-bit ResNet-18 achieves $\sim$6x speedups on NVIDIA T4 (less than 0.2ms per image). Also, the latency of 4-bit ResNet-18 is $\sim$53ms on FPGA, and $\sim$600ms on mobile ARM CPU [1]. Though our FDDA introduces a generator, it is only used in the training process and no extra parameters and latency are introduced in the inference stage.

References

1. Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., Gu, S.: Brecq: Pushing the limit of post-training quantization by block reconstruction. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)