Post Training Mixed Precision Quantization of Neural Networks using First-Order Information (Supplementary Material: Appendix)

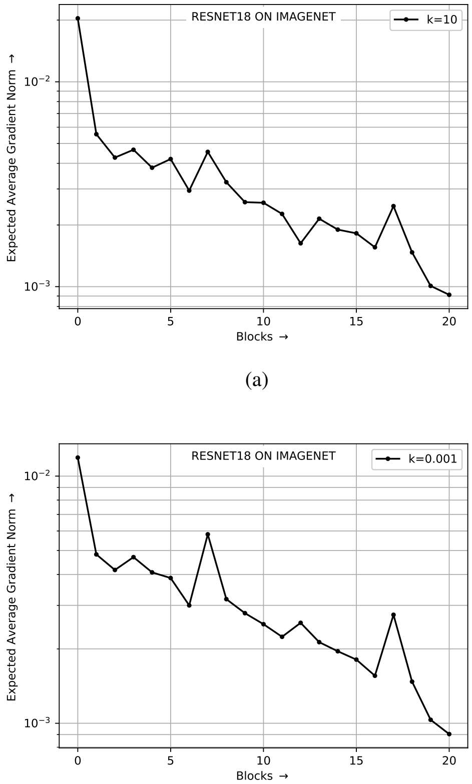
Arun Chauhan*, Utsav Tiwari*, Vikram N R
Samsung Research Institute
Bangalore, India
{arun.c, u.tiwari, vikram.nr}@samsung.com

1. Perturbation k for computing layers' sensitivity of the model

The magnitude of perturbation, k has a significant effect on computing expected average gradient norm of different blocks in the model. We performed extensive experiments for different values of k (perturbation) to compute the expected average gradient norm of all blocks in Resnet18 using the Algorithm 1 (please refer manuscript). For illustration, please refer Figure 1 which shows layers' sensitivity plot of all blocks in ResNet18 for two different values of k. Furthermore, to see the impact of k on accuracy of quantized model, we apply our methodology to quantize the ResNet18 for these two different values of k. Results shows that we get better accuracy of the quantized model when the perturbation is very close to the point of convergence of the pretrained model as shown in Table 1. For all the results in this work, we set k equals to 0.001 for computing expected average gradient norm.

Table 1: Quantization results of ResNet18 on ImageNet for different values of k (perturbation). The original model (FP32) accuracy of ResNet18 is 69.76. 'W Bits' and 'A Bits' stands for quantization bits used for weights and activations, respectively. The 'W CR' and 'A CR' stands for weight and activation compression ratio, respectively. The 'MP' refers to mixed-precision quantization, where we report the lowest bits used for weights and activations.

k	W	A	W	A	Top-1	Top-1
		Bit	CR	CR	Quant	Drop
10	2_{MP}	8	8 ×	4 ×	68.27	-1.49
0.001	2_{MP}	8	8 ×	4 ×	69.11	-0.65



(a)

(b)

Figure 1: Expected Average Gradient Norm of different blocks in ResNet18 on ImageNet for different values of perturbation k. It should be noted that the plot varies a lot for two different values of k. We provide the results for these two different settings on ResNet18 in Table 1.

*These authors contributed equally to this work.

Table 2: Comparison with state-of-the-art methods on ImageNet. 'RT' refers whether retraining of network is required or not . 'W Bits' and 'A Bits' stands for quantization bits used for weights and activations, respectively. The 'W CR' and 'A CR' stands for weight and activation compression ratio, respectively. The 'MP' refers to mixed-precision quantization, where we report the lowest bits used for weights and activations.

Network	Method	RT	Top-1 Full	W Bit	A Bit	W CR	A CR	Top-1 Quant	Top-1 Drop
ResNet-18	LQ-Nets* [7]	✓	70.30	3	32	6.10 ×	1.00 ×	69.30	-1.00
	ABC-Net [5]	✓	69.30	5	5	6.40 ×	6.40 ×	65.00	-4.30
	DoReFa* [8]	✓	70.40	5	5	5.16 ×	6.39 ×	68.40	-2.00
	PACT* [4]	✓	70.40	4	4	6.10 ×	7.98 ×	69.20	-1.20
	MPQNNCO [2]	✓	69.76	2_{MP}	8	10.66 ×	4.00 ×	69.39	-0.37
	DFQ [6]	✗	71.47	8	8	4 ×	4 ×	69.70	-1.77
	ZEROQ [1]	✗	71.47	8	8	4 ×	4 ×	71.43	-0.04
	DFPNMQ [3]	✗	69.76	MP	32	6.61 ×	1 ×	69.13	-0.63
	Ours	✗	**69.76**	2_{MP}	**8**	**4** ×	**4** ×	69.45	-0.31
	Ours	✗	**69.76**	2_{MP}	**8**	**8** ×	**4** ×	**69.11**	**-0.65**

* do not quantize the first and last layer

2. Sensitivity plots of different models under study

We use the Algorithm 1 (please refer manuscript) to compute the expected average gradient norm of different layers of MobileNet-V2 and Inception-V3 as shown in Figure 2. For ResNet18, please refer Figure 1b. We can clearly observe the comparable difference between the expected average gradient norm for different layers of the models.

3. Quantization results on ResNet18

For ResNet18, we compare the results with methods which do not require retraining (RT) after quantization and show that our method attains a smaller accuracy drop (-0.65%) with larger compression ratio (CR). We also achieve better performance with larger compression ratio in comparison to the methods which require retraining except MPQNNCO [2]. We achieve almost the original accuracy with 8× compression of weights with no training and require approximately 0.002% data which is a noticeable improvement against state of the art methods.

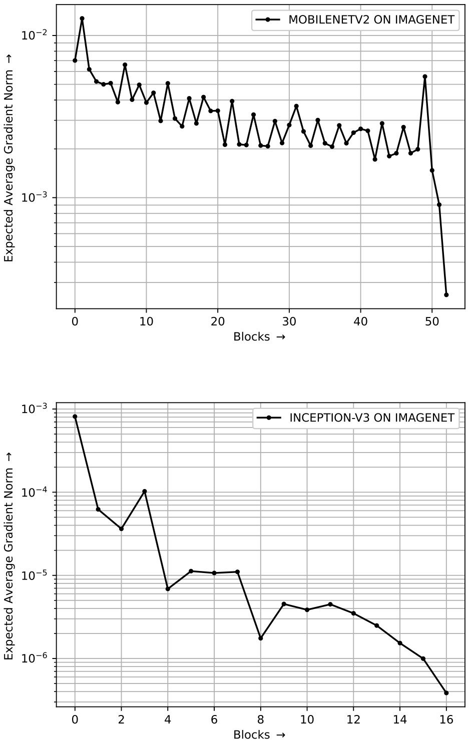


Figure 2: Expected Average Gradient Norm of different blocks in MobileNet-V2 *(top)* and Inception-V3 *(bottom)* on ImageNet.

References

[1] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13166–13175. Computer Vision Foundation / IEEE, 2020.

[2] Weihan Chen, Peisong Wang, and Jian Cheng. Towards mixed-precision quantization of neural networks via constrained optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages

5350–5359, October 2021.

[3] Vladimir Chikin and Mikhail Antiukh. Data-free network compression via parametric non-uniform mixed precision quantization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 450–459. IEEE, 2022.

[4] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. PACT: parameterized clipping activation for quantized neural networks. *CoRR*, abs/1805.06085, 2018.

[5] Xiaofan Lin, Cong Zhao, and Wei Pan. Towards accurate binary convolutional neural network. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 345–353, 2017.

[6] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1325–1334. IEEE, 2019.

[7] Dongqing Zhang, Jiaolong Yang, Dongqiangzi Ye, and Gang Hua. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, volume 11212 of *Lecture Notes in Computer Science*, pages 373–390. Springer, 2018.

[8] Shuchang Zhou, Zekun Ni, Xinyu Zhou, He Wen, Yuxin Wu, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *CoRR*, abs/1606.06160, 2016.