Generalizable Mixed-Precision Quantization via Attribution Rank Preservation

Ziwei Wang[1,2,3], Han Xiao[1,2,3], Jiwen Lu[1,2,3]*, Jie Zhou[1,2,3]
[1] Department of Automation, Tsinghua University, China
[2] State Key Lab of Intelligent Technologies and Systems, China
[3] Beijing National Research Center for Information Science and Technology, China
{wang-zw18, h-xiao20}@mails.tsinghua.edu.cn; {lujiwen,jzhou}@tsinghua.edu.cn

Appendix A: Visualization of the Optimal Quantization Policy

We searched the quantization policy on different small datasets with various architectures via the presented GMPQ. Figure 1 demonstrates the optimal bitwidth allocation for weights and activations of each layer, where ResNet18 was compressed and the policy was searched on various small datasets including CIFAR-10 [6], Cars [5], Flowers [8], Aircraft [7], Pets [9] and Food [1]. Figure 2 depicts the obtained quantization strategy searched on CIFAR-10 with MobileNet-V2 [10], ResNet18 [4] and ResNet50 architectures. The BOPs limit was set to 7.4G, 15.3G and 30.7G for MobileNet-V2, ResNet18 and ResNet50 compression respectively.

For quantization policy searched on different small datasets, the optimal bitwidth allocation varies significantly although the complexity of the obtained model is close to each other. It is observed that activations are usually assigned with higher bitwidth than weights in most quantization policy, indicating that the classification performance and attribution rank consistency are more sensitive to activation quantization than weight quantization. Meanwhile, the layers with residual connections usually have higher bitwidths for both weights and activations. The bitwidth distribution of weights and activations obtained on Cars, Aircraft, Food, and CIFAR-10 are similar, which also achieves better generalization performance on largescale datasets compared with that searched on Flowers and Pets. For the Flowers and Pets datasets, the optimal quantization policy is close to uniform quantization and the activations tend to receive lower bit allocation for quantization, which degrades the generalization ability of the mixed-precision networks.

For quantization policy for different architectures, it is observed that Layer 7, 12 and 17 in ResNet18 containing residual connections require the larger bitwidth compared with their corresponding regular branches. Since MobileNet-V2 is very compact, it receives higher bitwidth allocations than other network architectures. On the contrary, ResNet50 is compressed with lower bitwidth due to the significant redundancy compared with MobileNet-V2.

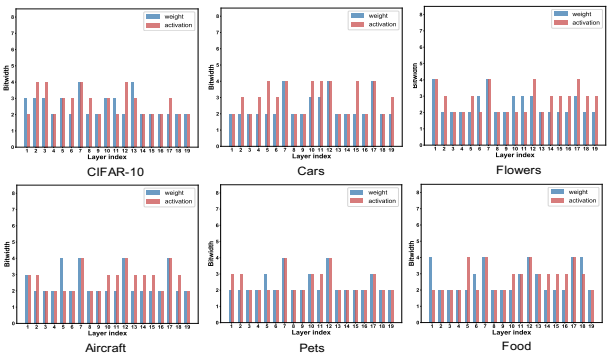

Figure 1. The visualization of the optimal quantization policy searched on different small datasets including CIFAR-10, Cars, Flowers, Aircraft, Pets and Food, where ResNet18 was compressed.

Appendix B: Accuracy of Quantization Policy Searched on Different Small Datasets

In this section, we show the top-1 accuracy and BOPs on ImageNet of our GMPQ with the quantization policy searched on different small datasets which include CIFAR-10, Cars, Flowers, Aircraft, Pets and Food. The applied network architectures contain MobileNet-V2, ResNet-18 and ResNet-50, and more accuracy-complexity tradeoffs for ResNet-18 is demonstrated in Figure 6(b) (main body). Table 1 illustrates the accuracy and the complexity on ImageNet, where those of full-precision networks are also provided. The search cost is significantly reduced across various architectures compared with conventional mixed-precision quantization methods shown in Table 1 (main body), while the accuracy is only degraded slightly. The accuracy of quantization policy searched on CIFAR-10 achieves the highest, because the gap of object

*Corresponding author

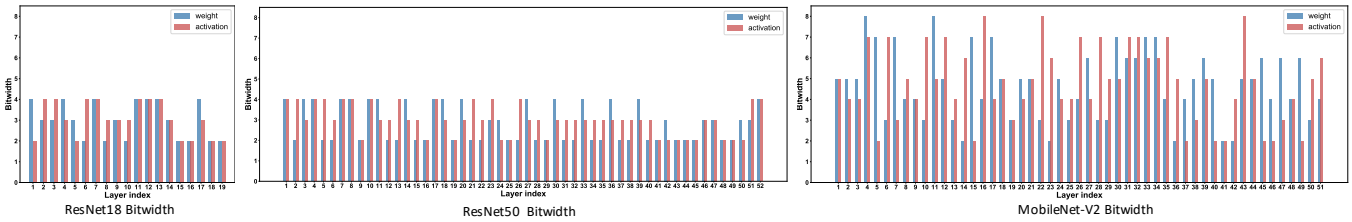

Figure 2. The visualization of the optimal quantization policy searched on CIFAR-10 by our GMPQ. We evaluated our method with MobileNet-V2, ResNet18 and ResNet50 on ImageNet for image classification.

Table 1. Top-1 accuracy (%) and BOPs (G) on ImageNet of the mixed-precision networks searched on different small datasets across various network architectures.

Architecture	Full-precision		CIFAR-10		Cars		Flowers		Aircraft		Pets		Food	
	Top1	BOPs	Top1	BOPs	Top1	BOPs	Top1	BOPs	Top1	BOPs	Top1	BOPs	Top1	BOPs
MobileNet-V2	71.9	337.9	70.4	7.4	69.8	7.2	67.8	7.9	69.9	7.5	66.7	7.8	69.9	7.1
ResNet18	69.7	1853.4	69.9	15.3	69.6	16.4	68.7	14.9	69.5	14.8	67.9	17.2	16.6	69.2
ResNet50	76.4	3952.6	75.8	30.7	75.5	29.8	73.8	33.2	75.6	29.5	73.3	34.1	75.6	32.7

category between CIFAR-10 and ImageNet is the smallest compared with other datasets. Although the discrepancy of object class distribution between ImageNet and the small datasets such as Aircraft is non-negligible, the accuracy of the mixed-precision networks is still comparable with state-of-the-art approaches shown in Table 1 (main body) due to the attribution rank preservation.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014.

[2] Zhaowei Cai and Nuno Vasconcelos. Rethinking differentiable search for mixed-precision neural networks. In *CVPR*, pages 2349–2358, 2020.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[5] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013.

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[7] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[8] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[9] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012.

[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.