Supplementary Materials of Instance-Aware Dynamic Neural Network Quantization

Zhenhua Liu[1,2], Yunhe Wang[2], Kai Han[2], Siwei Ma[1,3], Wen Gao[1,3]
[1]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[2] Huawei Noah's Ark Lab [3]Peng Cheng Laboratory
{liu-zh,swma,wgao}@pku.edu.cn, {yunhe.wang,kai.han}@huawei.com

1. Appendix

1.1. ResNet18 on ImageNet

We have shown the experimental results of ResNet18, which consists of a convolutional layer followed by 8 ResNet basic blocks. The results are shown in Table 1. DQNet can achieve 0.42% and 0.52% average top-1 accuracy gain over DoReFa and PACT.

Table 1. Comparison on the performance of proposed DQNet with PACT and DoReFa of ResNet18 on ImageNet dataset.

Method	Bit	Bit-FLOPs(G)	Top-1 Accuracy(%)
DoReFa	3	15.25	67.5
DoReFa	4	27.11	68.1
DoReFa	5	42.35	68.4
Average	–	28.24	68.00
DoReFa	∼3 MP	15.31	67.92
DoReFa	∼4 MP	27.19	68.41
DoReFa	∼5 MP	42.68	68.94
Average	–	28.39	**68.42**
PACT	3	15.25	68.1
PACT	4	27.11	69.2
PACT	5	42.35	69.8
Average	–	28.24	69.03
PACT	∼3 MP	15.36	68.49
PACT	∼4 MP	27.18	69.76
PACT	∼5 MP	42.49	70.40
Average	–	28.34	**69.55**

1.2. Effectiveness of bit-controller

We have shown the detailed bit-widths distribution of each layer for ResNet-20 on the validation set of CIFAR-10 in Figure 1. It is obvious that the bit controller produces various bit-widths for different samples, which demonstrates the effectiveness of the proposed bit-controller.

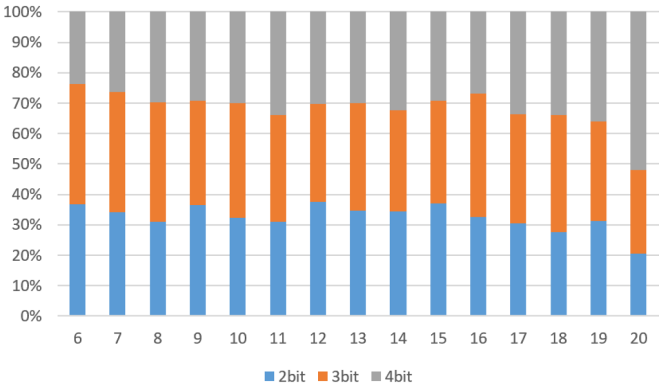

Figure 1. The bit-width distribution of each layer in ResNet-20 on CIFAR-10 dataset.