Supplementary Material of "PalQuant: Accelerating High-precision Networks on Low-precision Accelerators"

Qinghao Hu*[1], Gang Li*[2], Qiman Wu*[3], and Jian Cheng[1]✉

[1] Institute of Automation, Chinese Academy of Sciences
[2] Shanghai Jiao Tong University
[3] Baidu Inc.
huqinghao2014@ia.ac.cn, gliaca@sjtu.edu.cn, wuqiman@baidu.com, jcheng@nlpr.ia.ac.cn

1 Improve ShuffleNet v2 with Cyclic Shuffle

As mentioned in the main-body of the paper, the proposed cyclic shuffle shows its well performance on group convoltuions for quantized ResNet-18 and ResNet-34. Since we assume that it can serve as a complement of channel shuffle, here we explore the performance of cyclic shuffle on ShuffleNet v2.

As we didn't find the official training codes of ShuffleNet v2, we re-implemented the ShuffleNet v2 training algorithm under Pytorch framework. We adopt an SGD optimizer with the momentum set to 0.9. The weight decay is set to 1e-4, and the learning rate initialized by 0.1 is adjusted with a cosine learning rate decay strategy. All models in this subsection are trained from scratch for 300 epochs with a batch size of 512.

As shown in table 1, we can achieve +0.48% higher top-1 accuracy by adding one cyclic shuffle module at the beginning of stage3 and stage4 in ShuffleNet v2. This indicates that our proposed method can be applied on the deep neural networks that contains group convolutions or channel splitting modules to obtain higher accuracy.

Table 1. ShuffleNet V2 with Cyclic Shuffle

Module	Top1 Acc.	Top5 Acc.
ShuffleNet v2(our impl.)	68.71	88.48
+cyclic shuffle	**69.19** (+0.48)	**88.65** (+0.17)

* Equal Contribution.

2 BitOps Definition

Generally speaking, the number of floating-point operations (FLOPS) is the mainstream computational complexity metric. In Section 5 of the paper, we use the number of bit operations (BitOps)[1] to measure the computational complexity of quantized deep neural networks. For a convolutional layer with t kernels of size $c * k * k$, let h and w be the height and width of the output feature map respectively. Then the number of bit operations (BitOps) is:

$$\#\text{BitOps} = \mathrm{b_w} \times \mathrm{b_a} \times \mathrm{t} \times \mathrm{c} \times \mathrm{k} \times \mathrm{k} \times \mathrm{h} \times \mathrm{w} \tag{1}$$

where b_w and b_a denotes the bit-with of quantized weights and activations, respectively.

Table 2. Quantization Results of Plain-18 on ImageNet.

Method	Precision	BitOps	Top1 Acc.
Baseline	FP32	-	69.96
LSQ	4b A,4b W	29.10G	69.90
PalQuant	**B=2,G=2**	**14.87G**	**70.12**

3 PalQuant on CNNs without residual connections

To demonstrate the general applicability of PalQuant, we conduct experiments on Plain-18 network which has the same net architecture as ResNet-18 except for residual connections. Here we re-implement the baseline Plain-18 and LSQ method under the same training settings as PalQuant. Table 2 shows that PalQuant can achieve **+0.22**% higher top1 accuracy than LSQ with nearly only a half of BitOps. This result means that PalQuant can also be applied to deep networks without residual connections. As such, we think PalQuant can't be seen as an extension of ShuffleNet. Besides, PalQuant is a quantization method that aims to deploy high-precision networks on low-precision accelerators. One key component of PalQuant is expanding feature map channels and dividing them into groups. And the proposed cyclic shuffle is another contribution of PalQuant. These two contributions make PalQuant differ a lot from ShuffleNet.

References

1. Cai, Z., Vasconcelos, N.: Rethinking differentiable search for mixed-precision neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2349–2358 (2020)