Supplementary for Interleaved Group Convolutions

Ting Zhang[1] Guo-Jun Qi[2] Bin Xiao[1] Jingdong Wang[1]
[1]Microsoft Research [2]University of Central Florida
{tinzhan, Bin.Xiao, jingdw}@microsoft.com guojun.qi@ucf.edu

Appendices

Comparison with alternative structures. The proposed network is a stack of interleaved group convolution (IGC) blocks, where secondary group convolutions to blend the channels across partitions outputted by primary group convolutions.

In the main paper (Extensions and variants, Section 4), we discuss that a point-wise convolution, i.e., a 1×1 convolution, as an alternative of secondary group convolutions, introduces extra parameters and computation complexity. Such an alternative is also mentioned or discussed in Xception [1] and deep roots [2]. We denote this alternative block as Group-and-Point-wise Convolution (GPC). The number of parameters for one GPC block is $L \cdot M \cdot M \cdot S + L \cdot M \cdot L \cdot M$, where L is the number of partitions in group convolution, M is the number of channels in each partition, and thus LM is the number of total channels.

We replace IGC blocks in our networks using GPC blocks, with almost the same number of parameters. Table 1 presents the configurations: (i) #params ≈ 4672 and (ii) #params ≈ 17536 (See Table 1 in the main paper for the configurations of our IGC blocks). The results, together with our approach (with two secondary partitions) are presented in Figure 1 (More results about our networks are shown in Figure 3 in the main paper). We can see that our networks perform better and achieve around 0.5% improvement in both cases.

More on the connection to regular convolutions. We rewrite Equation (17) in the paper as the following,

$$\bar{\mathbf{x}}' = \mathbf{P}\mathbf{W}^d\mathbf{P}^\top\mathbf{W}^p\bar{\mathbf{x}}. \tag{1}$$

We discuss an extreme case: primary group convolution is a channel-wise convolution. Let

$$\bar{\mathbf{x}} = [\mathbf{x}^\top \mathbf{x}^\top \ \dots \ \mathbf{x}^\top]^\top.$$

be formed by concatenating $\mathbf{x}$ C times. The primary group convolution is given as follows,

$$\begin{aligned}\mathbf{W}^p = \mathrm{diag}(&\mathbf{w}_{11}^\top, \mathbf{w}_{12}^\top, \dots, \mathbf{w}_{1C}^\top, &(2)\\ &\mathbf{w}_{21}^\top, \mathbf{w}_{22}^\top, \dots, \mathbf{w}_{2C}^\top, &(3)\\ &\dots, &(4)\\ &\mathbf{w}_{C1}^\top, \mathbf{w}_{C2}^\top, \dots, \mathbf{w}_{CC}^\top), &(5)\end{aligned}$$

where $\mathbf{w}_{ij}$ is a vector of 3×3. The secondary group convolution is given as follows,

$$\mathbf{W}^d = \begin{bmatrix} \mathbf{1}^\top & \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{0}^\top \\ \mathbf{0}^\top & \mathbf{1}^\top & \mathbf{0}^\top & \mathbf{0}^\top \\ \dots & \dots & \dots & \dots \\ \mathbf{0}^\top & \mathbf{0}^\top & \dots & \mathbf{1}^\top \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}^\top & \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{0}^\top \\ \mathbf{0}^\top & \mathbf{1}^\top & \mathbf{0}^\top & \mathbf{0}^\top \\ \dots & \dots & \dots & \dots \\ \mathbf{0}^\top & \mathbf{0}^\top & \dots & \mathbf{1}^\top \end{bmatrix}, \tag{6}$$

where $\mathbf{1}$ is a vector of C ones, $\mathbf{0}$ is a vector of C zeros. $\mathbf{W}^d$ consists of $C \times 1$ blocks, with each block being a matrix of $C \times C^2$. Thus, we have

$$\bar{\mathbf{x}} = [\mathbf{x}'^\top \mathbf{x}'^\top \ \dots \ \mathbf{x}'^\top]^\top.$$

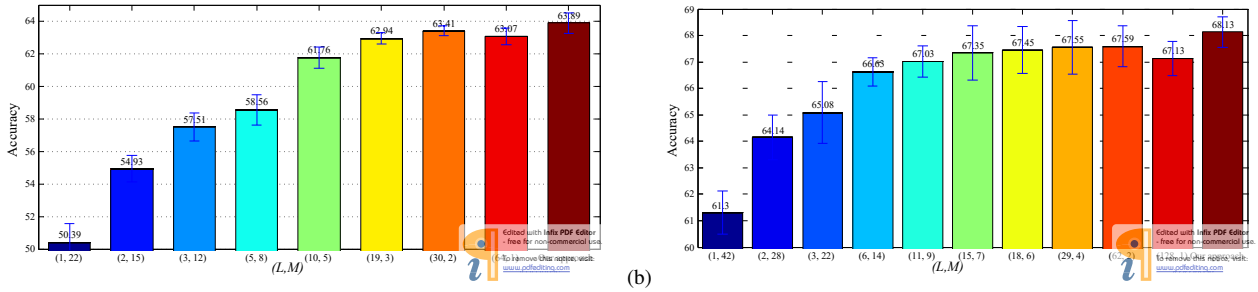
References

[1] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[2] Y. Ioannou, D. P. Robertson, R. Cipolla, and A. Criminisi. Deep roots: Improving CNN efficiency with hierarchical filter groups. *CoRR*, abs/1605.06489, 2016.

Table 1. Example configurations of GPCs for various numbers (L) of partitions and various numbers (M) of channels in each partition, under the roughly-equal number of parameters: (i) $\approx$ 4672 and (ii) $\approx$ 17536. The kernel size S in group convolutions is $9 = 3 \times 3$.

	(i) #params $\approx$ 4672									(ii) #params $\approx$ 17536										
	GPC								IGC	GPC										IGC
L	1	2	3	5	10	19	30	64	40	1	2	3	6	11	15	18	29	62	128	85
M	22	15	12	8	5	3	2	1	2	42	28	22	14	9	7	6	4	2	1	2
#params	4840	4950	5184	4480	4750	4788	4680	4672	4640	17640	17248	17424	17820	17640	17496	17632	17640	17608	17532	17510
Width	22	30	36	40	50	54	60	64	80	42	56	66	84	99	105	108	116	124	128	170



(a) (b)

Figure 1. Illustrating the performances between our approach and the networks stacking GPC with various numbers L and M with same #params on CIFAR-100. We report the mean and the standard deviation over five runs. (a) corresponds to (i) in Table 1 and (b) corresponds to (ii) with more parameters.