Similarity-Preserving Knowledge Distillation: Supplementary Material

Frederick Tung[1,2] and Greg Mori[1,2]
[1]Simon Fraser University [2]Borealis AI
ftung@sfu.ca, mori@cs.sfu.ca

Output size	MobileNet-k
112×112	$3 \times 3, 32k$
112×112	3×3 dw, $32k$ $1 \times 1, 64k$
56×56	3×3 dw, $64k$ $1 \times 1, 128k$
56×56	3×3 dw, $128k$ $1 \times 1, 128k$
28×28	3×3 dw, $128k$ $1 \times 1, 256k$
28×28	3×3 dw, $256k$ $1 \times 1, 256k$
14×14	3×3 dw, $256k$ $1 \times 1, 512k$
14×14	$\left.\begin{array}{c} 3 \times 3 \text{ dw}, 512k \\ 1 \times 1, 512k \end{array}\right\} \times 5$
7×7	3×3 dw, $512k$ $1 \times 1, 1024k$
7×7	3×3 dw, $1024k$ $1 \times 1, 1024k$
1×1	average pool, 47-d fc, softmax

Table 1. Structure of MobileNet networks used in transfer learning experiments. 'dw' denotes depthwise convolution. Downsampling is performed by strided 3×3 depthwise convolutions.

Output size	MobileNetV2-k
112×112	$3 \times 3, 32k$
112×112	bottleneck($t = 1, c = 16k, n = 1$)
56×56	bottleneck($t = 6, c = 24k, n = 2$)
28×28	bottleneck($t = 6, c = 32k, n = 3$)
14×14	bottleneck($t = 6, c = 64k, n = 4$)
14×14	bottleneck($t = 6, c = 96k, n = 3$)
7×7	bottleneck($t = 6, c = 160k, n = 3$)
7×7	bottleneck($t = 6, c = 320k, n = 1$)
7×7	$1 \times 1, 1280k$
1×1	average pool, 47-d fc, softmax

Table 2. Structure of MobileNetV2 networks used in transfer learning experiments. The notation 'bottleneck(t, c, n)' denotes a group of bottleneck residual blocks with expansion factor t, c output channels, and n repeated blocks. Downsampling is performed by strided 3×3 depthwise convolution in the first block of a group.

Output size	ShuffleNetV2-0.5	ShuffleNetV2-1.0	ShuffleNetV2-2.0
32×32	$3 \times 3, 24$	$3 \times 3, 24$	$3 \times 3, 24$
16×16	stage($c = 48, n = 4$)	stage($c = 116, n = 4$)	stage($c = 244, n = 4$)
8×8	stage($c = 96, n = 8$)	stage($c = 232, n = 8$)	stage($c = 488, n = 8$)
4×4	stage($c = 192, n = 4$)	stage($c = 464, n = 4$)	stage($c = 976, n = 4$)
4×4	$1 \times 1, 1024$	$1 \times 1, 1024$	$1 \times 1, 2048$
1×1	average pool, 10-d fc, softmax		

Table 3. Structure of ShuffleNetV2 networks used in CINIC-10 experiments. The notation 'stage(c, n)' denotes a group of ShuffleNetV2 building blocks with c output channels and n repeated blocks. Downsampling is performed by strided 3×3 depthwise convolutions in the first block of a group.