DMCP: Differentiable Markov Channel Pruning for Neural Networks Supplementary Material

Shaopeng Guo Yujie Wang Quanquan Li Junjie Yan
SenseTime Research
{guoshaopeng, wangyujie, liquanquan, yanjunjie}@sensetime.com

A. Structure of the Pruned Model

In this section, we provide the structure of our pruned models in various FLOPs settings. All pruned structures are sampled by the Expected Sampling method. In Table 7, we list the pruned structures of ResNet50 with 2.8G FLOPs. In Table 8, we list the pruned structures of MobileNetV2 with various FLOPs settings.

block	2.8G		
	conv1	conv2	conv3
-	51	-	-
block 1-1	38	42	223
block 1-2	47	47	
block 1-3	47	47	
block 2-1	103	101	461
block 2-2	92	96	
block 2-3	93	95	
block 2-4	100	100	
block 3-1	218	215	945
block 3-2	200	212	
block 3-3	208	213	
block 3-4	209	213	
block 3-5	213	215	
block 3-6	212	215	
block 4-1	459	454	1735
block 4-2	455	455	
block 4-3	457	437	

Table 7. The number of channels in each layer of the pruned ResNet50, the “block” column indicates the index of residual blocks. In each block, we only list the output channel of each layer. The first block marked by “-” is conv1 layer, whose input channel is 3. The fully connected layer is omitted because its output channel is fixed to 1000 for ImageNet classification.

B. Visualization

B.1. FLOPs Distribution of the Pruned Model

We sample 3000 structures from the trained MobileNetV2-210M via the Markov process, and the distribution of their FLOPs is showed in Figure 6. From the figure, we can find that the mean of FLOPs lies around 210M, which means that the expected FLOPs converged to the desired budget 210M.

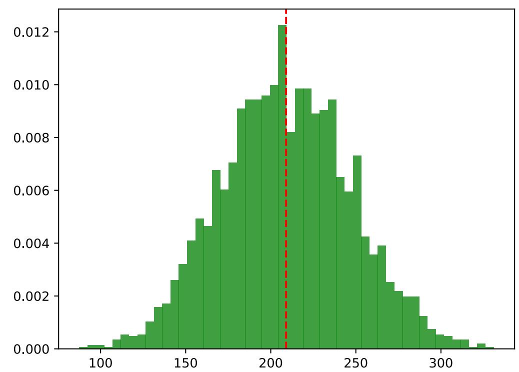

Figure 6. The FLOPs distribution of 3000 structures sampled from MobileNetV2-210M by Markov process. The x-axis is the MFLOPs and the y-axis is the frequency. The red dashed line is the mean of the FLOPs of 3000 sampled structures. The FLOPs of the unpruned network is 672M.

B.2. The Channel Distribution of Pruned Layers.

In this section, we examine the channel distribution in each layer of the pruned model. We sample 3000 models from MobileNetV2-210M whose FLOPs are within the desired budget (210M) by Markov process. Figure 7 shows the channel distribution of 12 layers sampled from different blocks. From the figure, we can observe that the number of channel in most layers follows an uni-modal distribution, and some layers choose to retain all the channels (e.g. LinearBottleneck6 and 7).

C. Comparison between using warm-up and using pre-trained model

As described in Section 3.2, the warm-up is performed by only running stage 1 that updating the weights of the unpruned network by our proposed variant sandwich rule,

block	operation	300M		210M		97M		59M	
		input	output	input	output	input	output	input	output
-	conv1	3	15	3	13	3	8	3	6
bottleneck 1-1	conv2	15	15	13	13	8	8	6	6
	conv3	15	11	13	11	8	7	6	10
bottleneck 2-1	conv1	11	51	11	45	7	33	10	18
	conv3	51	19	45	19	33	10	18	12
bottleneck 2-2	conv1	19	57	19	63	10	32	12	16
	conv3	57	19	63	19	32	10	16	12
bottleneck 3-1	conv1	19	126	19	110	10	60	16	32
	conv3	126	34	110	30	60	17	32	15
bottleneck 3-2	conv1	34	105	30	118	17	59	15	41
	conv3	105	34	118	30	59	17	41	15
bottleneck 3-3	conv1	34	109	30	113	17	59	15	41
	conv3	109	34	113	30	59	17	41	15
bottleneck 4-1	conv1	34	246	30	223	17	154	15	98
	conv3	246	79	223	64	154	46	98	36
bottleneck 4-2	conv1	79	267	64	241	46	156	36	120
	conv3	267	79	241	64	156	46	120	36
bottleneck 4-3	conv1	79	291	64	256	46	194	36	120
	conv3	291	79	256	64	194	46	120	36
bottleneck 4-4	conv1	79	284	64	272	46	156	36	120
	conv3	284	79	272	64	156	46	120	36
bottleneck 5-1	conv1	79	486	64	415	46	270	36	158
	conv3	486	102	415	80	270	60	158	45
bottleneck 5-2	conv1	102	384	80	337	60	177	45	123
	conv3	384	102	337	80	177	60	123	45
bottleneck 5-3	conv1	102	422	80	361	60	576	45	123
	conv3	422	102	361	80	576	60	123	45
bottleneck 6-1	conv1	102	775	80	694	60	462	45	351
	conv3	775	231	694	191	462	144	351	96
bottleneck 6-2	conv1	231	980	191	858	144	576	96	480
	conv3	980	231	858	191	576	144	480	96
bottleneck 6-3	conv1	231	1082	191	933	144	672	96	576
	conv3	1082	231	933	191	672	144	576	96
bottleneck 7-1	conv1	231	1411	191	1283	144	864	96	768
	conv3	1411	417	1283	262	864	192	768	128

Table 8. The number of channels in pruned MobileNetV2 in various FLOPs settings. The "block" column shows different linear bottlenecks. We only list the number of output channels of conv1 and conv3 in each block because conv2 is depth-wise convolution and its number of channels is equal to the output channels of conv1.

which makes the channel group more important than the channel group right after it, providing a good initialization for iterative training. However, using a pre-trained model cannot provide initialization with the property. Our experiment also shows the superiority of using warm-up, on DMCP-MBV2 with 210M FLOPs by replacing warm-up with using pre-trained models, a 0.6% accuracy drop was observed.

D. Modeling architecture parameters as independent Bernoulli variables

Given a layer with C channels, the solution space of our method is $O(C)$, by modeling architecture parameters as Bernoulli variables, the solution space becomes $O(2^C)$, as there are 2^C possible channel combinations, which makes it much harder to optimize. To demonstrate our analysis, we experiment on MobileNet-v2 with 210M FLOPs, by replacing Markov modeling with Bernoulli Modelling of architecture parameters, the performance of the pruned model is 70.1%, which is 2.3% lower than DMCP.

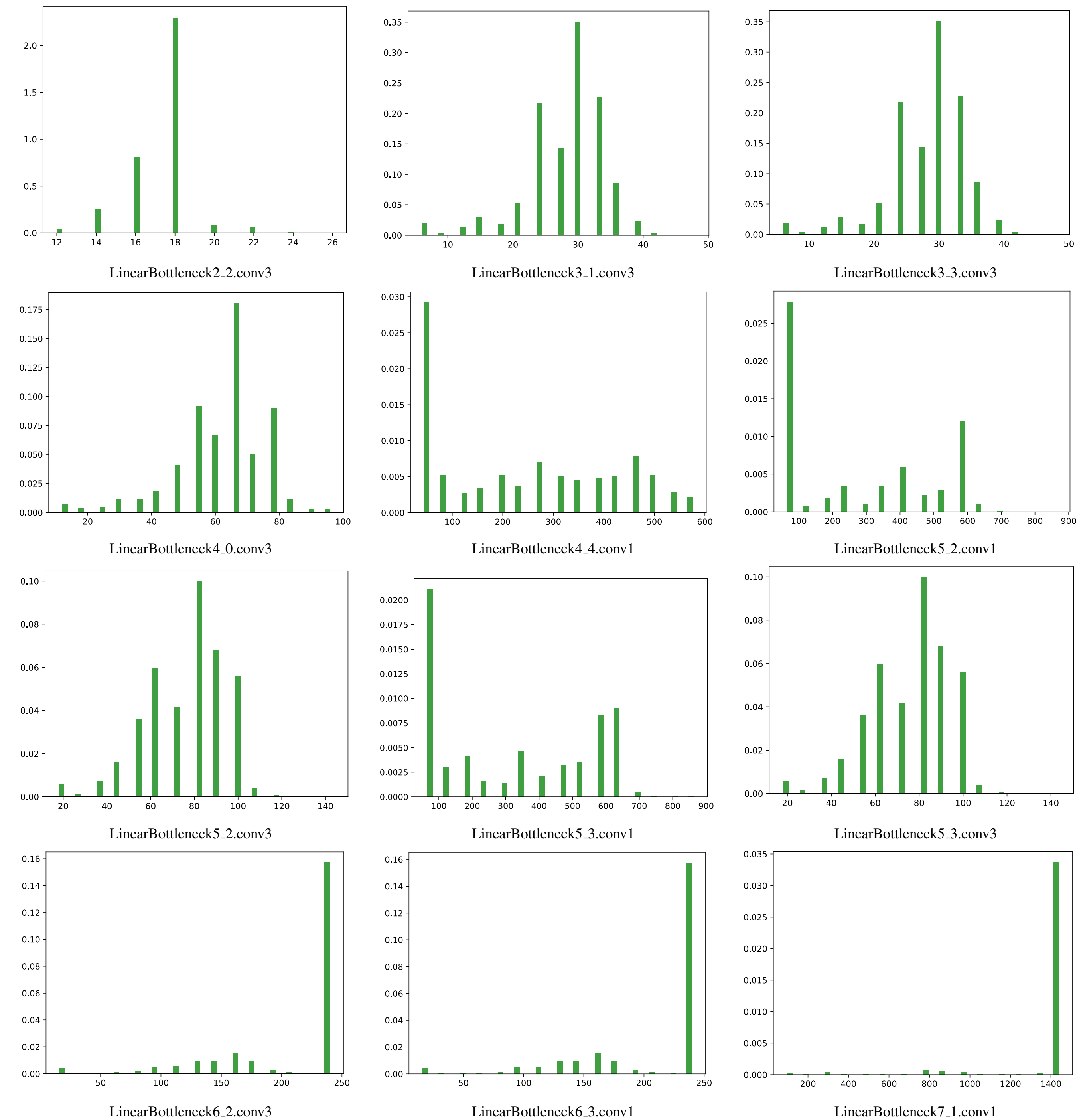

Figure 7. The channel distribution of 12 layers sampled from different blocks in MobileNetV2-210M. The y-axis is the frequency and the x-axis is the number of channels. Note that we divided channels each layer into 15 groups.