A. Experimental details

Here we explain our experimental setup. For all architectures, we optimize our network by minimizing cross-entropy loss using SGD.

A.1. MobileNetV1+BFT

We have used weight decay of 10^{-5}. We train for 170 epochs. We have used a constant learning rate 0.5 and decay it by $\frac{1}{10}$ at epochs 140, 160. For details on width multiplier of MobileNet and input resolution on each experiment look at Table 4.

A.2. ShuffleNetV2+BFT

We have used weight decay of 10^{-5}. We train for 300 epochs. We start with a learning rate of 0.5 linearly decaying it to 0. All of the pointwise convolutions are replaced by BFT as shown in Figure 6, except the first pointwise convolution with input channel size of 24. For comparing under the similar number of FLOPs we have slightly changed ShuffleNet's layer width to create ShuffleNetV2-1.25. This is the structure which is used for shuffleNetV2-1.25:

Layer	output size	Kernel	Stride	Repeat	Width
Image	224×224				3
Conv1 Max pool	112×112 56×56	3×3 3×3	2 2	1	24
Stage 2	28×28 28×28		2 1	1 3	128
Stage 3	14×14 14×14		2 1	1 7	256
Stage 4	7×7 7×7		2 1	1 3	1024
Conv 5	7×7	BFT	1	1	1024
Global Pool	1×1	7×7			
FC					1000
FLOPS					41

For details on input resolution on each experiment look at Table 5.

A.3. MobileNetV3+BFT

We have used weight decay of 10^{-5}. We train for 200 epochs. We start with a warm-up for the first 5 epochs, starting from a learning rate 0.1 and linearly increasing it to 0.5. Then we decay learning rate from 0.5 to 0.0 using a cosine scheme in the remaining 195 epochs. For details on width multiplier and input resolution on each experiment look at Table 6.

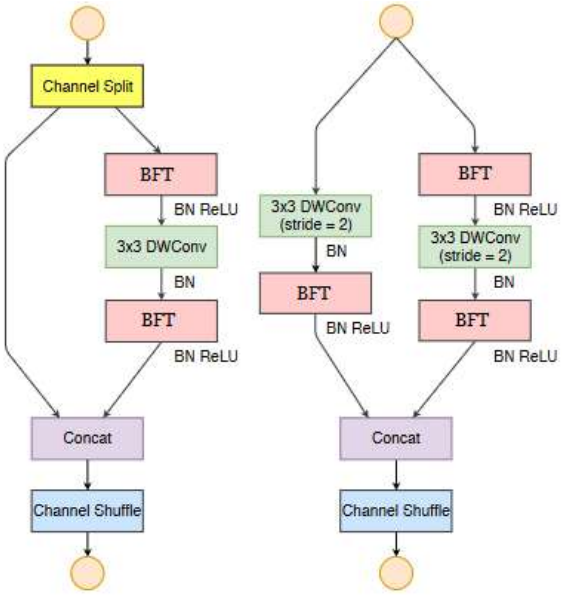

Figure 6:
ShuffleNetV2+BFT Block

.

MobileNet				MobileNet+**BFT**				gain
width	resolution	flops	Accuracy	width	resolution	flops	Accuracy	
0.25	128	14 M	41.50	1.0	96	14 M	46.58	**5.08**
0.25	160	21 M	45.50	1.0	128	23 M	52.26	**6.76**
0.25	192 224	34 M 41 M	47.70 50.60	1.0	160	35 M	54.30	**6.60** **3.70**
0.50	128	49 M	56.30	1.0 2.0	192 128	51 M 52 M	57.56 58.35	**1.26** **2.05**
0.50	192	110 M	61.70	2.0	192	112 M	63.03	**1.33**
0.50	224	150 M	63.30	2.0	224	150 M	64.32	**1.02**

Table 4: Comparision between Mobilenet and Mobilenet+BFT. For comparision under similar number of FLOPs we have used wider channels in MobileNet+BFT.

ShuffleNetV2				ShuffleNetV2+**BFT**				gain
width	resolution	flops	Accuracy	width	resolution	flops	Accuracy	
0.50	128	14 M	50.86*	1.25	128	14 M	55.26	**4.4**
0.50	160	21 M	55.21*	1.25	160	21 M	57.83	**2.62**
0.50	224	41 M	59.70* 60.30	1.25	224	41 M	61.33	**1.63** **1.03**

Table 5: Comparision between ShuffleNetV2 and ShuffleNetV2+BFT. For comparision under similar number of FLOPs we have used wider channels in ShuffleNetV2+BFT.

MobileNetV3				MobileNetV3+**BFT**				gain
width	resolution	flops	Accuracy	width	resolution	flops	Accuracy	
Small-0.35	224	13 M	49.8	Small-0.5	224	15 M	55.21	**5.41**

Table 6: Comparision between MobileNetV3 and MobileNetV3+BFT. For comparision under similar number of FLOPs we have used wider channels in MobileNetV3+BFT.