Appendix for Cache and Reuse: Rethinking the Efficiency of On-device Transfer Learning

8. Additional Experimental Results for Table 2 (Section 5.2)

Table 7 presents more experimental results for Table 2 in Section 5.2. These results further prove the strength of our method. Indeed, our two-stage on-device transfer learning method can significantly boost the training speed, while providing a similar or even better accuracy.

Model	Type	Latency [ms]	Data [MB]	Accuracy [%]					
				CF10	CF100	Cars	Flowers	Food	Pets
MobileNet-V2 Last 1 Block	Baseline	3.30	664	91.07	70.53	**65.59**	91.45	74.27	**87.79**
	Ours-8b	**0.50**	**375**	91.41	72.07	64.15	**92.00**	74.39	87.71
	Ours-4b	**0.50**	**188**	**91.67**	**72.60**	64.59	91.82	74.42	87.54
	Ours-2b	**0.50**	**95**	90.71	70.01	59.76	91.32	**75.12**	87.71
	Ours-1b	**0.50**	**48**	88.79	66.77	51.20	89.36	71.43	86.97
MobileNet-V2 Last 4 Blocks	Baseline	3.87	664	94.21	76.28	75.79	93.02	77.52	87.52
	Ours-8b	**1.34**	898	94.49	76.35	76.87	93.49	78.50	87.76
	Ours-4b	**1.33**	**450**	**94.60**	**76.60**	**77.63**	**93.49**	**78.86**	87.60
	Ours-2b	**1.34**	**226**	94.43	76.33	77.15	93.40	79.14	**87.98**
	Ours-1b	**1.35**	**113**	92.34	69.85	72.78	90.13	75.81	86.59
MobileNet-V2 Last 7 Blocks	Baseline	4.70	664	**95.10**	**77.74**	80.13	92.41	79.43	**87.52**
	Ours-8b	**2.47**	**599**	94.86	76.94	80.50	**92.57**	79.79	87.38
	Ours-4b	**2.49**	**300**	94.83	77.18	**80.67**	92.42	80.10	87.38
	Ours-2b	**2.49**	**151**	94.84	76.99	80.49	91.72	**80.17**	87.30
	Ours-1b	**2.48**	**76**	93.40	70.83	75.39	88.60	76.66	85.91
EfficientNet-B0 Last 3 Blocks	Baseline	7.41	664	94.19	76.49	74.58	**93.46**	**79.62**	91.50
	Ours-8b	**2.34**	**450**	94.50	76.86	74.80	92.76	78.82	91.36
	Ours-4b	**2.32**	**226**	**94.45**	**77.09**	**75.20**	92.78	79.15	91.44
	Ours-2b	**2.34**	**113**	94.38	76.76	74.17	92.39	79.24	**91.55**
	Ours-1b	**2.33**	**57**	93.54	74.29	67.09	89.72	76.81	91.47
EfficientNet-B0 Last 5 Blocks	Baseline	8.66	664	95.10	78.65	79.21	**94.57**	**81.59**	**91.09**
	Ours-8b	**3.94**	1048	95.47	**79.18**	**80.60**	92.94	81.01	90.84
	Ours-4b	**3.94**	**525**	95.54	79.12	80.05	92.88	81.14	90.71
	Ours-2b	**3.94**	**263**	**95.82**	78.83	80.55	93.71	81.04	90.60
	Ours-1b	**3.96**	**132**	94.27	74.06	74.37	89.88	76.02	89.97
ResNet-18 Last 2 Blocks	Baseline	2.87	664	93.63	75.47	74.98	**92.03**	75.84	**88.85**
	Ours-8b	**1.16**	2393	93.79	76.43	**75.13**	91.84	**75.79**	88.53
	Ours-4b	**1.15**	1200	93.82	**76.38**	74.85	91.93	75.84	88.50
	Ours-2b	**1.14**	**599**	**93.83**	76.24	74.18	92.01	75.27	88.72
	Ours-1b	**1.13**	**300**	92.71	74.21	69.69	90.01	74.26	88.74
EfficientFormerV2-S0 Last 5 Blocks (ViT)	Baseline	6.62	664	94.20	**74.97**	**77.58**	**91.51**	**81.98**	**90.57**
	Ours-8b	**2.60**	898	94.07	73.55	74.73	90.40	79.65	88.83
	Ours-4b	**2.61**	**450**	94.10	73.95	75.35	90.34	80.24	89.07
	Ours-2b	**2.61**	**226**	**94.21**	74.22	75.60	90.83	80.72	89.15
	Ours-1b	**2.65**	**113**	93.40	70.95	65.99	86.66	75.61	87.65

Table 7. Additional experimental results for Table 2. Ours-1/2/4/8b refers to the proposed method with 2/4-bit quantization for cache compression. For the baseline, metric "Data" represent the size of the dataset, and for our two-stage training, metric "Data" represents the size of the cache, as shown in Figure 2. Latency and data size are measured by training with CIFAR100. For latency and data size, we highlight results that outperform the baseline; for accuracy, we highlight the highest accuracy for each dataset.