A. Experiment setups

Our experiments are done on the CIFAR-10 dataset [17] and the ImageNet ILSVRC-2012 dataset [28]. We access both datasets via the API provided in the "TorchVision" Python package. As recommended in the PyTorch tutorial, we normalize the data and augment the data with random crop and random horizontal flip before the training. We use batch size 100 to train CIFAR-10 model and use 256 for the ImageNet model. For all the models on CIFAR-10, both the full-rank SVD training and the low-rank finetuning are trained for 164 epochs. The learning rate is set to 0.001 initially and decayed by 0.1 at epoch 81 and 122. For models on ImageNet, the full-rank SVD training is trained for 90 epochs, with initial learning rate 0.1 and learning rate decayed by 0.1 every 30 epochs. The low-rank finetuning is done for 60 epochs, starting at learning rate 0.01 and decay by 0.1 at epoch 30. We use pretrained full-rank decomposed model (trained with the orthogonality regularizer but without sparsity-inducing regularizer) to initialize the SVD training. SGD optimizer with momentum 0.9 is used for optimizing all the models, with weight decay 5e-4 for CIFAR-10 models and 1e-4 for ImageNet models. The accuracy reported in the experiment is the best validation accuracy achieved during the finetuning process.

During the SVD training, the decay parameter of the orthogonality regularizer λ_o is set to 1.0 for both channel-wise and spatial-wise decomposition on CIFAR-10. On ImageNet, for training the ResNet-18 model λ_o is set to 5.0 for both decomposition methods. For the ResNet-50 model, λ_o is set to 10.0 for channel-wise decomposition and 5.0 for spatial-wise decomposition. The decay parameter λ_s for the sparsity-inducing regularizer and the energy threshold used for singular value pruning are altered through different set of experiments to fully explore the accuracy-#FLOPs tradeoff. In most cases, the energy threshold is selected through a line search, where we find the highest percentage of energy that can be pruned without leading to a sudden accuracy drop. The λ_s and the energy thresholds used in each set of the experiments are reported alongside the experiment results in Appendix B.

B. Detailed experiment results

In this section we list the exact data used to plot the experiment result figures in Section 4. The results of our proposed method with various choice of decomposition method and sparsity-inducing regularizer tested on the CIFAR-10 dataset are listed in Table 3. All of these data points are visualized in Figure 2 and Figure 3 to compare the tradeoff tendency under different conditions. As discussed in Section 4.5, the results of spatial-wise decomposition with the Hoyer regularizer for ResNet-20 and ResNet-32 are shown in Figure 4 to compare with previous methods. The results of both channel-wise and spatial-wise decomposition with the Hoyer regularizer are compared with previous methods in Figure 4 for ResNet-56 and ResNet-110. For experiments on ImageNet dataset, the results of our method for the ResNet-18 model are listed in Table 4, and the results of our method for the ResNet-50 model are listed in Table 5.

The baseline results of previous works on compressing CIFAR-10 and ImageNet models used for comparison in Figure 4 are listed in Table 6-8. As there are a large amount of previous works in this field, we only list the results of the most recent works here to show the state-of-the-art Pareto frontier. Therefore we choose state of the art low-rank compression methods like Jaderberg et al. [15], Zhang et al. [39], TRP [35] and C-SGD [5], as well as recent filter pruning methods like NISP [37], SFP [10] and CNN-FCF [20] as the baseline to compare our results against.

Model	Reg Type	Decay	Energy Pruned	Accuracy Gain(%)	Speed Up
ResNet-20	Hoyer	0.03	1.5e-5	0.04	2.20 ×
Channel		0.07	6.0e-6	-0.27	2.66 ×
		0.1	3.0e-6	-0.54	2.94 ×
Base Acc:	L1	0.01	7.0e-2	1.13	1.43 ×
90.93%		0.001	2.7e-2	0.63	1.59 ×
		0.1	1.0e-1	0.32	2.10 ×
		0.3	1.0e-1	-0.48	2.84 ×
	None	0.0	1.9e-1	-0.37	2.03 ×
		0.0	2.8e-1	-0.52	2.54 ×
		0.0	3.3e-1	-0.61	2.88 ×
ResNet-20	Hoyer	0.01	1.0e-3	0.40	3.26 ×
Spatial		0.03	2.0e-5	-0.10	3.87 ×
		0.1	4.0e-6	-0.86	4.77 ×
Base Acc:		0.01	7.0e-3	-1.03	5.16 ×
90.99%	L1	0.01	6.0e-2	0.58	2.26 ×
		0.1	1.0e-1	-0.52	3.55 ×
		0.3	1.0e-1	-0.83	4.79 ×
	None	0.0	2.9e-1	0.59	2.44 ×
		0.0	3.9e-1	-0.78	3.15 ×
		0.0	4.8e-1	-1.30	4.05 ×
ResNet-32	Hoyer	0.003	3.0e-3	0.04	2.22 ×
Channel		0.01	1.0e-4	-0.10	2.44 ×
		0.03	2.0e-6	-0.86	2.56 ×
Base Acc:	L1	0.03	2.0e-2	0.22	1.58 ×
92.12%		0.1	1.0e-1	-0.21	2.84 ×
		0.3	5.0e-2	-0.96	3.08 ×
	None	0.0	1.8e-1	-0.30	2.23 ×
		0.0	2.1e-1	-0.11	2.41 ×
		0.0	2.3e-1	-0.27	2.51 ×
ResNet-32	Hoyer	0.001	5.0e-2	0.52	2.56 ×
Spatial		0.005	5.0e-3	-0.38	3.93 ×
		0.01	8.0e-4	-0.62	4.57 ×
Base Acc:		0.03	8.0e-6	-1.12	5.30 ×
92.14%	L1	0.03	7.0e-2	0.13	2.60 ×
		0.1	2.5e-2	-0.34	4.20 ×
		0.1	1.5e-1	-0.96	5.32 ×
	None	0.0	3.8e-1	-0.60	3.62 ×
		0.0	4.8e-1	-1.76	4.71 ×
		0.0	5.3e-1	-2.14	5.34 ×
ResNet-56	Hoyer	0.001	2.0e-2	0.39	2.70 ×
Channel		0.003	1.0e-3	-0.29	3.49 ×
		0.01	7.0e-6	-0.41	4.35 ×
Base Acc:		0.01	2.0e-5	-0.68	4.94 ×
93.28%		0.03	3.0e-7	-1.20	5.16 ×
	L1	0.1	3.0e-2	-0.30	4.25 ×
		0.1	1.5e-1	-0.59	4.86 ×
	None	0.0	2.8e-1	-0.16	2.91 ×
		0.0	3.8e-1	-0.98	3.71 ×
		0.0	4.7e-1	-1.78	4.70 ×
ResNet-56	Hoyer	0.001	3.0e-2	0.17	3.07 ×
Spatial		0.003	1.0e-3	-0.09	3.75 ×

		0.01	1.0e-4	-0.70	5.43 ×
Base Acc:		0.03	1.0e-6	-1.37	6.90 ×
93.36%	L1	0.03	5.0e-3	-0.24	3.19 ×
		0.03	5.0e-2	-0.90	5.61 ×
		0.03	2.5e-1	-1.38	6.76 ×
	None	0.0	2.8e-1	-0.18	2.96 ×
		0.0	4.7e-1	-0.47	4.76 ×
		0.0	5.2e-1	-2.22	5.43 ×
ResNet-110	Hoyer	0.001	5.0e-3	0.38	3.85 ×
Channel		0.003	3.0e-4	-0.34	5.00 ×
		0.01	3.0e-7	-0.60	6.66 ×
Base Acc:		0.03	1.0e-6	-1.27	8.76 ×
93.58%	L1	0.03	1.0e-1	-0.28	5.02 ×
		0.03	3.0e-1	-1.27	7.44 ×
	None	0.0	3.7e-1	-0.32	4.26 ×
		0.0	4.6e-1	-1.86	5.44 ×
		0.0	5.5e-1	-2.59	7.03 ×
ResNet-110	Hoyer	0.001	1.3e-2	0.10	4.75 ×
Spatial		0.003	7.0e-4	-0.46	6.42 ×
		0.01	2.0e-5	-1.28	8.76 ×
Base Acc:		0.03	2.0e-8	-2.03	10.06×
93.93%	L1	0.03	3.0e-2	-0.42	5.02 ×
		0.03	1.0e-1	-0.67	6.45 ×
		0.03	1.5e-1	-1.01	7.21 ×
		0.03	2.5e-1	-1.36	8.66 ×
	None	0.0	4.7e-1	-1.56	5.69 ×
		0.0	5.6e-1	-2.27	7.55 ×
		0.0	6.1e-1	-3.44	8.87 ×

Table 3: Full results of applying the proposed method on ResNet models on the CIFAR-10 dataset with various hyperparameters. [Decay] marks the decay variable for the sparse regularization, i.e. λ_s. [Energy Pruned] means the energy threshold used for singular value pruning, i.e. e. [Accuracy Gain] denotes the gain of Top-1 accuracy from the accuracy of the baseline full-rank model. [Speed Up] is computed as the ratio of #FLOPs of the original model and the achieved low-rank model.

Table 4. Results of applying the proposed method on ResNet-18 model on the ImageNet dataset. Hoyer regularizer is used as the sparsity-inducing regularizer for the singular values. Top-5 validation accuracy is reported in the [Base Acc] and the [Accuracy Gain] columns. [Speed Up] is computed as the ratio of #FLOPs of the original model and the achieved low-rank model.

Decompose	Base Acc	Decay	Energy Pruned	Accuracy Gain	Speed Up
Channel	88.54%	0.002	5.0e-4	0.94%	1.45 ×
		0.003	1.0e-4	-1.28%	2.03 ×
		0.005	1.0e-4	-2.47%	2.98 ×
		0.01	1.0e-5	-4.20%	4.21 ×
Spatial	88.54%	0.002	1.0e-4	0.67%	1.61 ×
		0.005	1.0e-4	-0.84%	2.98 ×
		0.01	1.0e-4	-3.13%	6.36 ×

Table 5. Results of applying the proposed method on ResNet-50 model on the ImageNet dataset. Hoyer regularizer is used as the sparsity-inducing regularizer for the singular values. Top-5 validation accuracy is reported in the [Base Acc] and the [Accuracy Gain] columns. [Speed Up] is computed as the ratio of #FLOPs of the original model and the achieved low-rank model.

Decompose	Base Acc	Decay	Energy Pruned	Accuracy Gain	Speed Up
Channel	91.72%	0.001	1.0e-4	0.02%	1.37 ×
		0.002	1.0e-4	-0.12%	1.92 ×
		0.003	5.0e-5	-0.54%	2.51 ×
		0.005	5.0e-5	-1.56%	4.17 ×
Spatial	91.91%	0.0005	1.0e-3	0.06%	1.44 ×
		0.001	1.0e-4	-0.10%	1.79 ×
		0.002	2.0e-4	-1.09%	3.05 ×

Table 6. Baselines on the CIFAR-10 dataset. [Accu.↑] means the Top-1 accuracy gain comparing to that of the full model. [Sp. Up] denotes speed up computed as the ratio of #FLOPs before and after the model compression. [-] is marked when no result is available in the paper.

Method	ResNet-20		ResNet-32		ResNet-56		ResNet-110	
	Accu.↑	Sp. Up	Accu.↑	Sp. Up	Accu.↑	Sp. Up	Accu.↑	Sp. Up
Zhang et al.	-3.61%	1.41 ×	-2.76%	1.41 ×	-	-	-	-
Jaderberg et al.	-2.25%	1.66 ×	-2.29%	1.68 ×	-	-	-	-
TRP-Ch	-0.43%	2.17 ×	-0.72%	2.20 ×	-	-	-	-
TRP-Sp	-0.37%	2.84 ×	-0.75%	3.40 ×	-	-	-	-
SFP	-1.37%	1.79 ×	-0.55%	1.71 ×	0.19%	1.70 ×	0.18%	1.69 ×
CNN-FCF	-1.07%	1.71 ×	-0.25%	1.73 ×	0.24%	1.75 ×	0.09%	1.76 ×
	-2.67%	3.17 ×	-1.69%	3.36 ×	-1.22%	3.44 ×	-0.62%	2.55 ×
C-SGD-5/8	-	-	-	-	0.23%	2.55 ×	0.03%	2.56 ×
Nisp	-	-	-	-	-0.03%	1.77 ×	-0.18%	1.78 ×

Table 7. Baselines of compressing ResNet-18 model on the ImageNet dataset. [Accu.↑] means the Top-5 accuracy gain comparing to that of the full model. [Sp. Up] denotes speed up computed as the ratio of #FLOPs before and after the model compression.

Channel-wise			Spatial-wise		
Method	Accu.↑	Sp. Up	Method	Accu.↑	Sp. Up
Zhang et al.	-4.85%	1.39×	Jaderberg et al.	-4.82%	2.00×
Zhang et al.	-4.10%	1.41×	TRP-Sp	-1.80%	2.60×
TRP-Ch	-2.06%	1.81×	TRP-Sp	-2.71%	3.20×
TRP-Ch	-2.91%	2.20×	TRP-Sp	-3.24%	3.68×
TRP-Ch	-3.02%	2.50×			

Table 8. Baselines of compressing ResNet-50 model on the ImageNet dataset. [Accu.↑] means the Top-5 accuracy gain comparing to that of the full model. [Sp. Up] denotes speed up computed as the ratio of #FLOPs before and after the model compression.

Method	Accu.↑	Sp. Up	Method	Accu.↑	Sp. Up
SFP	-0.81%	1.72×	NISP-50-A	-0.21%	1.38×
CNN-FCF-A	+0.26%	1.41×	NISP-50-B	-0.89%	1.79×
CNN-FCF-B	-0.19%	1.85×	C-SGD-70	-0.10%	1.58×
CNN-FCF-C	-0.69%	2.33×	C-SGD-50	-0.29%	1.86×
CNN-FCF-D	-1.37%	2.96×	C-SGD-30	-0.47%	2.26×