Supplementary Material for "Group Sparsity: The Hinge Between Filter Pruning and Decomposition for Network Compression"

Yawei Li[1], Shuhang Gu[1], Christoph Mayer[1], Luc Van Gool[1,2], Radu Timofte[1]
[1]Computer Vision Lab, ETH Zürich, Switzerland, [2]KU Leuven, Belgium
{yawei.li, shuhang.gu, chmayer, vangool, radu.timofte}@vision.ee.ethz.ch

1. Closed-form Solutions to the Proximal Operators

The proximal operator of a given function $f(\cdot)$ is defined by

$$\mathbf{prox}_{\lambda f} = \arg\min_{v} \left\{ f(v) - \frac{1}{\lambda}\|x - v\|_2^2 \right\} \tag{1}$$

This operator has closed-form solution when the function $f(\cdot)$ has the form of ℓ_1, $\ell_{1/2}$, ℓ_{1-2}, and logsum regularization. For ℓ_1, the solution is the soft-thresholding function and for $\ell_{1/2}$ it is the so-called half-thresholding function. The soft-thresholding function is defined as

$$\mathcal{S}_\lambda(x) = \mathrm{sgn}(x)[|x| - \lambda]_+, \tag{2}$$

where $\mathrm{sgn}(\cdot)$ is the sign function and $[\cdot]_+$ calculates the maximum of the argument and 0. The hard-thresholding function is given by

$$\mathcal{H}_\lambda(x) = \begin{cases} \frac{2}{3}x(1 + \cos(\frac{2\pi}{3} - \frac{2}{3}\phi_\lambda(x))), & |x| > \frac{\sqrt[3]{54}}{4}(\lambda)^{\frac{2}{3}}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $\phi_\lambda(x) = \arccos(\frac{\lambda}{8}(\frac{|x|}{3})^{-\frac{3}{2}})$.

The $\ell_{2,1}$ group sparsity regularizer is defined as

$$\mathcal{R}(\mathbf{A}) = \Phi(\|\mathbf{A}_g\|_2) = \sum_g \|\mathbf{A}_g\|_2^p, \tag{4}$$

where $\Phi(\cdot)$ is the function of the group ℓ_2 norms $\|\mathbf{A}_g\|_2$ and has the form of ℓ_1 norm here. The proximal operator of the sparsity-inducing matrix $\mathbf{A}$ defined in the main paper is

$$\mathbf{A}_{t+1} = \mathbf{prox}_{\lambda\eta\mathcal{R}}(\mathbf{A}_{t+\Delta}) = \arg\min_{\mathbf{A}} \left\{ \mathcal{R}(\mathbf{A}_{t+\Delta}) + \frac{1}{2\lambda\eta}\|\mathbf{A} - \mathbf{A}_{t+\Delta}\|_F^2 \right\}, \tag{5}$$

where the function $\mathcal{R}(\cdot)$ replaces $f(\cdot)$ in Eqn. 1. The closed-form solution of the proximal operator in Eqn. 5 can be derived from the solutions to Eqn. 1 according to the following theorem [1].

Theorem 1 *Let $f : \mathbb{E} \to \mathbb{R}$ be a function given by $f(\mathbf{x}) = g(\|\mathbf{x}\|)$, where $g : \mathbb{R} \to (-\infty, \infty]$ is a proper closed and convex function satisfying $\mathrm{dom}(g) \subseteq [0, \infty)$. Then,*

$$\mathbf{prox}_{\lambda f}(\mathbf{x}) = \begin{cases} \mathbf{prox}_{\lambda g}(\|\mathbf{x}\|_2)\frac{\mathbf{x}}{\|\mathbf{x}\|_2}, & \mathbf{x} \neq \mathbf{0}, \\ \{\mathbf{u} \in \mathbb{E} : \|\mathbf{u}\|_2 = \mathbf{prox}_{\lambda g}(\mathbf{0})\}, & \mathbf{x} = \mathbf{0}. \end{cases} \tag{6}$$

Thus, with a little bit variable substitution, when $\Phi(\cdot)$ is ℓ_1 regularizer, the solution to Eqn. 5 is given by

$$\mathbf{A}_{t+1} = \left[1 - \frac{\lambda\eta}{\|\mathbf{A}_g\|_2}\right]_+ \mathbf{A}_{g,i}, \tag{7}$$

Regularizer	Solution
ℓ_1	$\mathbf{A}_{t+1} = \left[1 - \frac{\lambda\eta}{\|\mathbf{A}_g\|_2}\right]_+ \mathbf{A}_{g,i}$
$\ell_{1/2}$	$\mathbf{A}_{t+1} = \begin{cases} \frac{2}{3}\left(1 + \cos\left(\frac{2\pi}{3} - \frac{2}{3}\phi_{\lambda\eta}\left(\|\mathbf{A}_g\|_2\right)\right)\right)\mathbf{A}_{g,i}, & \|\mathbf{A}_g\|_2 > \frac{\sqrt[3]{54}}{4}(\lambda\eta)^{\frac{2}{3}}, \\ 0, & \text{otherwise}, \end{cases}$ $\phi_{\lambda\eta}(\|\mathbf{A}_g\|_2) = \arccos(\frac{\lambda\eta}{8}(\frac{\|\mathbf{A}_g\|_2}{3})^{-\frac{3}{2}})$
ℓ_{1-2}	$\mathbf{A}_{t+1} = (1 + \frac{\lambda\eta}{\|\mathbf{c}\|_2})[1 - \frac{\lambda\eta}{\|\mathbf{A}_g\|_2}]_+\mathbf{A}_{g,i}$ $\mathbf{c}_g = [\|\mathbf{A}_g\|_2 - \lambda\eta]_+$
logsum	$\mathbf{A}_{t+1} = \begin{cases} \frac{c_1+\sqrt{c_2}}{2}\frac{\mathbf{A}_{g,i}}{\|\mathbf{A}_g\|_2}, & c_2 > 0, \\ 0, & c_2 \leqslant 0, \end{cases}$ $\lambda > 0, 0 < \epsilon < \sqrt{\lambda\eta}, c_1 = \|\mathbf{A}_g\|_2 - \epsilon, c_2 = c_1^2 - 4(\lambda\eta - \epsilon\|\mathbf{A}_g\|_2)$

Table 1: The solution to the proximal operator for ℓ_1, ℓ_{1-2}, $\ell_{1/2}$, and logsum regularizers.

Regularizer	ℓ_1	ℓ_{1-2}	$\ell_{1/2}$	logsum
Regularization factor λ	$2e^{-4}$	$2e^{-4}$	$4e^{-4}$	$9e^{-5}$

Table 2: The regularization factor for ℓ_1, ℓ_{1-2}, $\ell_{1/2}$, and logsum regularizers.

where $\mathbf{A}_{g,i}$ is the i-th element in the g-th group of the sparsity-inducing matrix $\mathbf{A}$, and for the sake of simplicity, the subscript $_{t+\Delta}$ is omitted.

When the function $\Phi(\cdot)$ has the form of $\ell_{1/2}$, ℓ_{1-2}, and logsum, it is non-convex. However, we still use the variable substitution in Theorem 1 experimentally and the corresponding results in the main paper are also very competitive. For $\ell_{1/2}$ regularizer, the solution is given by

$$\mathbf{A}_{t+1} = \begin{cases} \frac{2}{3}\left(1 + \cos\left(\frac{2\pi}{3} - \frac{2}{3}\phi_{\lambda\eta}\left(\|\mathbf{A}_g\|_2\right)\right)\right)\mathbf{A}_{g,i}, & \|\mathbf{A}_g\|_2 > \frac{\sqrt[3]{54}}{4}(\lambda\eta)^{\frac{2}{3}}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where $\phi_{\lambda\eta}(\|\mathbf{A}_g\|_2) = \arccos(\frac{\lambda\eta}{8}(\frac{\|\mathbf{A}_g\|_2}{3})^{-\frac{3}{2}})$. Similarly, the solution to the logsum regularizer is given by

$$\mathbf{A}_{t+1} = \begin{cases} \frac{c_1+\sqrt{c_2}}{2}\frac{\mathbf{A}_{g,i}}{\|\mathbf{A}_g\|_2}, & c_2 > 0, \\ 0, & c_2 \leqslant 0, \end{cases} \tag{9}$$

where $\lambda > 0$, $0 < \epsilon < \sqrt{\lambda\eta}$, $c_1 = \|\mathbf{A}_g\|_2 - \epsilon$, and $c_2 = c_1^2 - 4(\lambda\eta - \epsilon\|\mathbf{A}_g\|_2)$. When the regularizer is ℓ_{1-2} regularizer, then the solution is given by

$$\mathbf{A}_{t+1} = (1 + \frac{\lambda\eta}{\|\mathbf{c}\|_2})[1 - \frac{\lambda\eta}{\|\mathbf{A}_g\|_2}]_+\mathbf{A}_{g,i} \tag{10}$$

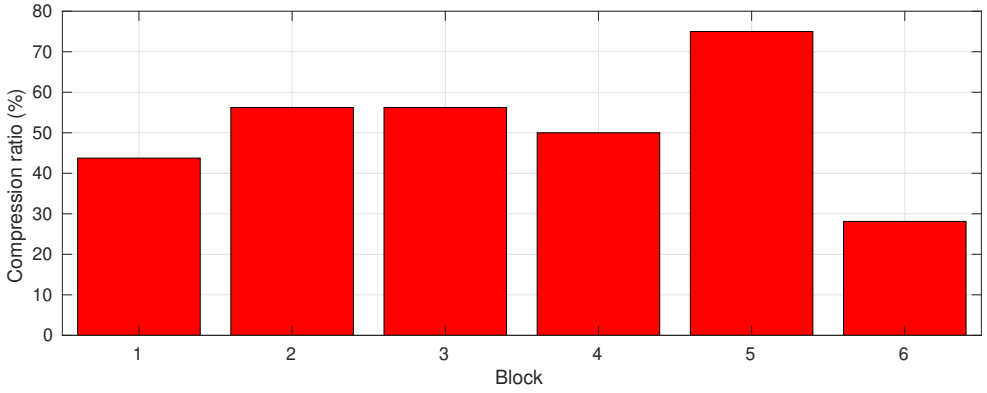
where $\mathbf{c}_g = [\|\mathbf{A}_g\|_2 - \lambda\eta]_+$. Note that the case where all of the group ℓ_2 norms $\mathbf{A}_g$ equal 0 is not considered [6] because it never happens during the optimization of our algorithm. The solutions are summarized in Table 1.

2. Hyper Parameters for Different Regularizers

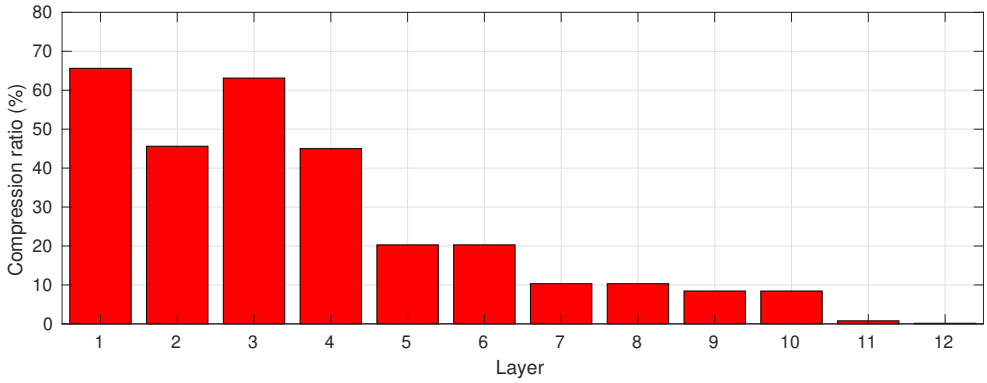
The regularization factors for different regularizers are listed in Table 2. For CIFAR10 and CIFAR100 datasets, the learning rate η of the sparsity-inducing matrix $\mathbf{A}$ during compression optimization is set to 0.1. The ratio between the learning rate of $\mathbf{W}$ and $\mathbf{A}$ is set to 0.01. That is, the learning rate η_s of $\mathbf{W}$ during compression optimization is 0.001. For ImageNet, both η and η_s during optimization are set to 0.001.

3. More Parameter Comparison

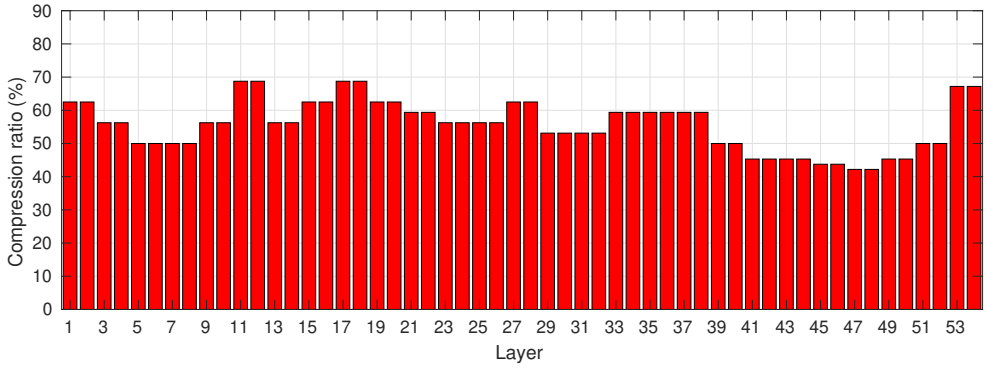
In Fig. 3 and Fig. 4, more parameter comparison results are shown. The figures report several operating points of the proposed method and SSS [4]. The proposed method forms a lower error bound for SSS. In Fig. 3, our Hinge method without



(a) ResNeXt20 [5], CIFAR100.



(b) WRN [7], CIFAR100.



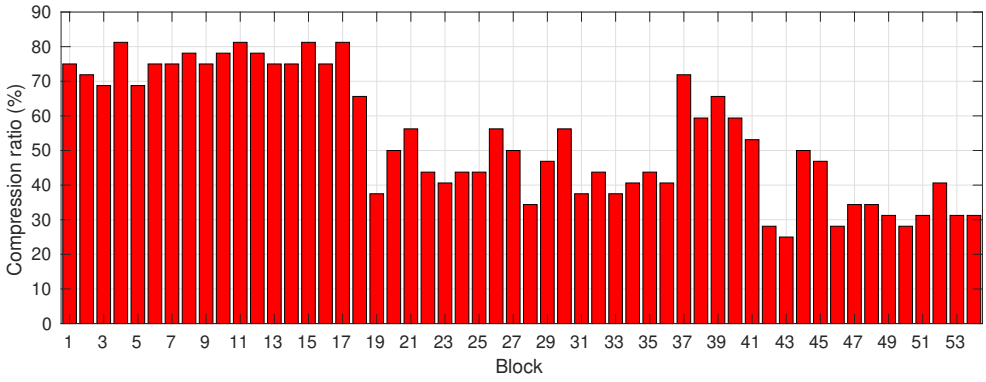
(c) ResNet56 [2], CIFAR10.

Figure 1: The layer-wise or block-wise compression ratio of the model resulting from the proposed method.

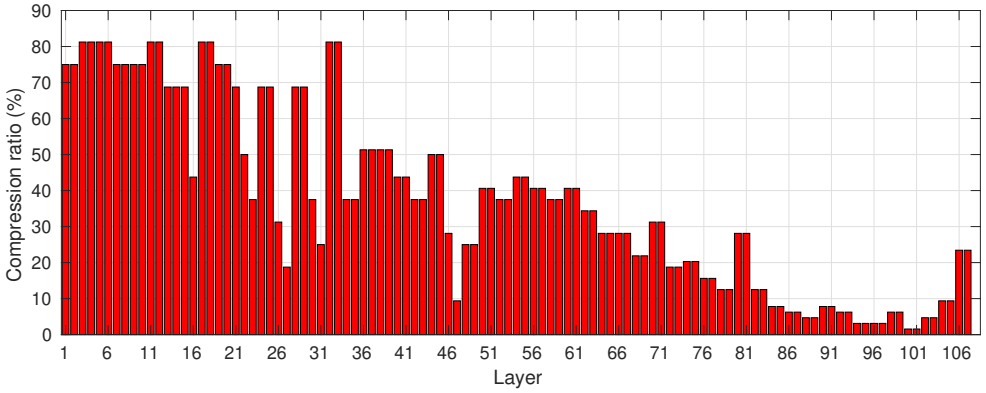
distillation loss is already better than SSS. And with the distillation loss, the proposed method shoots even lower Top-1 error rate.

4. Layer-wise Compression Ratio

The layer-wise or block-wise compression ratio of the model compressed by the proposed method is shown in Fig. 1 and Fig. 2, respectively. For ResNeXt [5], the aim is to compress the 3×3 convolution in the residual block and the two 1×1 convolutions are used as the sparsity-inducing matrices. Thus, the block-wise compression ratio is reported. For ResNet [2] and WRN [7], there are two 3×3 convolutions in each residual block. Each of the two convolutions is compressed by introducing a sparsity-inducing matrix. Thus, the layer-wise compression ratio is reported. As shown in the Fig. 1b, Fig. 2a

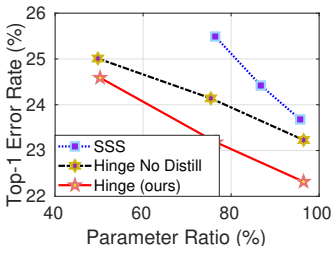


(a) ResNeXt164 [5].

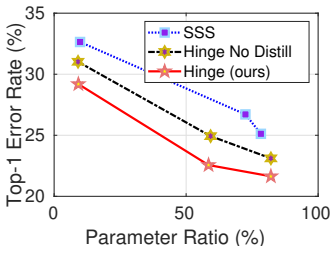


(b) ResNet164 [2].

Figure 2: The layer-wise or block-wise compression ratio of the model resulting from the proposed method. All results are reported for CIFAR100.



(a) ResNet164



(b) ResNeXt164

Figure 3: Comparison between SSS [4] and the proposed method. Top-1 error rate is reported for CIFAR100.

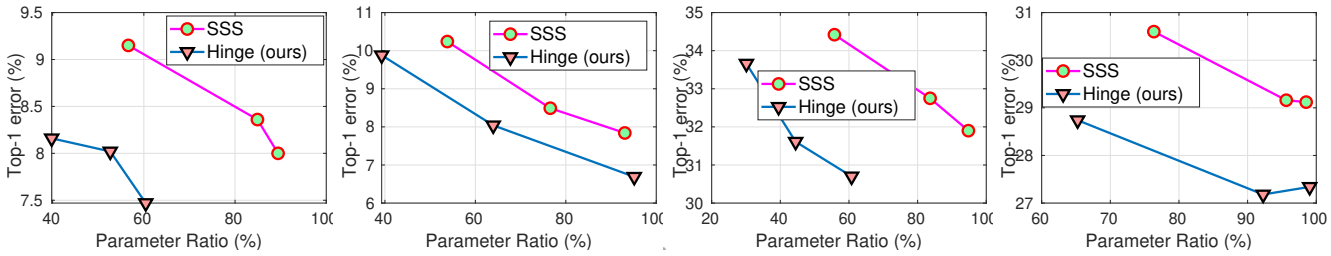
and Fig. 2b, for WRN, ResNeXt164, and ResNet164, our approach tends to compress the shallow layers more compared with the deep layers. This is consistent with former research [3]. As for ResNet56 in Fig 1c, the proposed method results in a sawtooth architecture. That is, for the convolutions with the same feature dimension (*i.e*. Layer 1 to Layer 18, Layer 19 to Layer 36, and Layer 37 to Layer54), the middle layers generally have a severer degree of compression.

References

[1] Amir Beck. *First-order methods in optimization*, volume 25. SIAM, 2017. 4321

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 4323, 4324

[3] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proc. ICCV*, pages 1389–1397, 2017. 4324



(a) CIFAR10, ResNet20 (b) CIFAR10, ResNeXt20 (c) CIFAR100, ResNet20 (d) CIFAR100, ResNeXt20

Figure 4: Comparison between SSS [4] and the proposed method. Top-1 error rate is reported. (a) and (b) shows the results on CIFAR10 while (c) and (d) shows the results on CIFAR100.

[4] Zehao Huang and Naiyan Wang. Data-driven sparse structure selection for deep neural networks. In *Proc. ECCV*, pages 304–320, 2018. 4322, 4324, 4325

[5] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. CVPR*, pages 1492–1500, 2017. 4323, 4324

[6] Quanming Yao, James T Kwok, and Xiawei Guo. Fast learning with nonconvex L1-2 regularization. *arXiv preprint arXiv:1610.09461*, 2016. 4322

[7] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. 2016. 4323