Supplementary Materials for Towards Compact CNNs via Collaborative Compression

Yuchao Li[1,2*], Shaohui Lin[3*], Jianzhuang Liu[4], Qixiang Ye[5], Mengdi Wang[2]
Fei Chao[1], Fan Yang[6], Jincheng Ma[6], Qi Tian[4], Rongrong Ji[1,7,8†]

[1]Media Analytics and Computing Laboratory, Department of Artificial Intelligence,
School of Informatics, Xiamen University, [2]Alibaba Group, [3]East China Normal University
[4]Huawei Noah's Ark Lab, [5]University of Chinese Academy of Sciences, [6]Huawei TechnologiesCo.,Ltd
[7]Institute of Artificial Intelligence, Xiamen University, [8]Peng Cheng Laboratory

{laiyin.lyc, didou.wmd}@alibaba-inc.com, shlin@cs.ecnu.edu.cn, qxye@ucas.ac.cn,
{liu.jianzhuang, yangfan74}@huawei.com, majincheng1@hisilicon.com, {fchao, rrji}@xmu.edu.cn

A. Derivation of Equation (14) in the Paper

The important metric in our method is:

$$P_o^{l,(t)} = I_o^{l,(t)} + \gamma \frac{1}{|\mathcal{U}^{l,(t)}|-1} \sum_{i \in \mathcal{U}^{l,(t)} \setminus o} I_{i|o}^{l,(t)}. \tag{1}$$

The second item in it can be re-formulated as:

$$\begin{aligned}
\gamma \frac{1}{|\mathcal{U}^{l,(t)}|-1} \sum_{i \in \mathcal{U}^{l,(t)} \setminus o} I_{i|o}^{l,(t)} &= \gamma \frac{1}{|\mathcal{U}_o^{l,(t)}|} \sum_{i \in \mathcal{U}_o^{l,(t)}} \mathrm{S}[(\mathcal{G}^l * (\overline{\mathcal{W}}_{i|o}^{l,(t)} - \mathcal{W}^l)^2] \\
&= \gamma \frac{1}{|\mathcal{U}_o^{l,(t)}|} \sum_{i \in \mathcal{U}_o^{l,(t)}} \mathrm{S}[(\mathcal{G}^l * (\overline{\mathcal{W}}_{i|o}^{l,(t)} - \overline{\mathcal{W}}_o^{l,(t)} + \overline{\mathcal{W}}_o^{l,(t)} - \mathcal{W}^l))^2]. \\
&= \gamma \frac{1}{|\mathcal{U}_o^{l,(t)}|} \sum_{i \in \mathcal{U}_o^{l,(t)}} \mathrm{S}[(\mathcal{G}^l * (\theta_{i|o}^{l,(t)} + \theta_o^{l,(t)}))^2] \\
&= \gamma \frac{1}{|\mathcal{U}_o^{l,(t)}|} \sum_{i \in \mathcal{U}_o^{l,(t)}} \mathrm{S}[(\mathcal{G}^l * \theta_{i|o}^{l,(t)})^2 + 2\mathcal{G}^l * \theta_{i|o}^{l,(t)} * \mathcal{G}^l * \theta_o^{l,(t)} + (\mathcal{G}^l * \theta_o^{l,(t)})^2] \\
&= \gamma \frac{1}{|\mathcal{U}_o^{l,(t)}|} \mathrm{S}[(\mathcal{G}^l)^2 * \sum_{i \in \mathcal{U}_o^{l,(t)}} (\theta_{i|o}^{l,(t)})^2] \\
&\quad + \gamma \frac{2}{|\mathcal{U}^{l,(t)}|} \mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \sum_{i \in \mathcal{U}^{l,(t)} \setminus o} \theta_{i|o}^{l,(t)}] + \gamma \mathrm{S}[(\mathcal{G}^l * \theta_o^{l,(t)})^2],
\end{aligned} \tag{2}$$

where $\theta_{i|o}^{l,(t)} = \overline{\mathcal{W}}_{i|o}^{l,(t)} - \overline{\mathcal{W}}_o^{l,(t)}$, $\theta_o^{l,(t)} = \overline{\mathcal{W}}_o^{l,(t)} - \mathcal{W}^l$. $*$ represents element-wise multiplication, and $\mathcal{U}_o^{l,(t)} = \mathcal{U}^{l,(t)} \setminus o$. Then, we consider the compression units in channel pruning $\mathcal{U}_{cp|o}^{l,(t)}$ and tensor decomposition $\mathcal{U}_{td|o}^{l,(t)}$ separately, where $\mathcal{U}_{cp|o}^{l,(t)} \cup \mathcal{U}_{td|o}^{l,(t)} = \mathcal{U}_o^{l,(t)}$. For the remaining compression units of channel pruning (*i.e.,* input channels), the first item in Eq. 2 can be rewritten

*Equal contribution.
†Corresponding author.

as:

$$\gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \sum_{i\in\mathcal{U}_{cp|o}^{l,(t)}}(\theta_{i|o}^{l,(t)})^2] = \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * (-\overline{\mathcal{W}}_o^{l,(t)})^2] = \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l * \overline{\mathcal{W}}_o^{l,(t)})^2], \tag{3}$$

and the second item can also be rewritten as:

$$\gamma\frac{2}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \sum_{i\in\mathcal{U}_{cp|o}^{l,(t)}}\theta_{i|o}^{l,(t)}] = -\gamma\frac{2}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \overline{\mathcal{W}}_o^{l,(t)}]. \tag{4}$$

Therefore, the second item in Eq. 1 for channel pruning becomes:

$$\gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\sum_{i\in\mathcal{U}_{cp|o}^{l,(t)}} I_{i|o}^{l,(t)} = \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l * \overline{\mathcal{W}}_o^{l,(t)})^2] - \gamma\frac{2}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \overline{\mathcal{W}}_o^{l,(t)}] + \gamma\frac{|\mathcal{U}_{cp|o}^{l,(t)}|}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l * \theta_o^{l,(t)})^2]. \tag{5}$$

For the remaining compression units of tensor decomposition (*i.e.,* singular values), the first item in Eq. 2 can be rewritten as:

$$\begin{aligned}
\gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \sum_{i\in\mathcal{U}_{td|o}^{l,(t)}}(\theta_{i|o}^{l,(t)})^2] &= \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \sum_{i\in\mathcal{U}_{td|o}^{l,(t)}}\phi(-U_{:,i|o}^{l,(t)}\Sigma_{i,i|o}^{l,(t)}{V_{i,:|o}^{l,(t)}}^\top)^2] \\
&= \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \sum_{i\in\mathcal{U}_{td|o}^{l,(t)}}\phi((U_{:,i|o}^{l,(t)})^2(\Sigma_{i,i|o}^{l,(t)})^2({V_{i,:|o}^{l,(t)}}^\top)^2)] \\
&= \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \phi((U_o^{l,(t)})^2(\Sigma_o^{l,(t)})^2({V_o^{l,(t)}}^\top)^2)],
\end{aligned} \tag{6}$$

where $\overline{\mathcal{W}}_o^{l,(t)} = \phi(U_o^{l,(t)}\overline{\Sigma}_o^{l,(t)}{V_o^{l,(t)}}^\top)$ is the SVD of $\phi^{-1}(\overline{\mathcal{W}}_o^{l,(t)})$. The second item of Eq.2 can also be rewritten as:

$$\begin{aligned}
\gamma\frac{2}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \sum_{i\in\mathcal{U}_{td|o}^{l,(t)}}\theta_{i|o}^{l,(t)}] &= \gamma\frac{2}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \sum_{i\in\mathcal{U}_{td|o}^{l,(t)}}\phi(-U_{:,i|o}^{l,(t)}\Sigma_{i,i|o}^{l,(t)}{V_{i,:|o}^{l,(t)}}^\top)] \\
&= -\gamma\frac{2}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \phi(U_o^{l,(t)}\Sigma_o^{l,(t)}{V_o^{l,(t)}}^\top)] \\
&= -\gamma\frac{2}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \overline{\mathcal{W}}_o^{l,(t)}].
\end{aligned} \tag{7}$$

Therefore, for the tensor decomposition, the second item in the importance metric Eq. 1 is:

$$\begin{aligned}
\gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\sum_{i\in\mathcal{U}_{td|o}^{l,(t)}} I_{i|o}^{l,(t)} =& \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \phi((U_o^{l,(t)})^2(\Sigma_o^{l,(t)})^2({V_o^{l,(t)}}^\top)^2)] \\
&- \gamma\frac{2}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \overline{\mathcal{W}}_o^{l,(t)}] + \gamma\frac{|\mathcal{U}_{td|o}^{l,(t)}|}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l * \theta_o^{l,(t)})^2].
\end{aligned} \tag{8}$$

Finally, the importance metric can be formulated as follows:

$$\begin{aligned}
P_o^{l,(t)} &= I_o^{l,(t)} + \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\sum_{i\in\mathcal{U}_{cp|o}^{l,(t)}} I_{i|o}^{l,(t)} + \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\sum_{i\in\mathcal{U}_{td|o}^{l,(t)}} I_{i|o}^{l,(t)} \\
&= (1+\gamma)\mathrm{S}[(\mathcal{G}^l * (\overline{\mathcal{W}}_o^{l,(t)} - \mathcal{W}^l))^2] - \gamma\frac{4}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \theta_o^{l,(t)} * \overline{\mathcal{W}}_o^{l,(t)}] \\
&+ \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l * \overline{\mathcal{W}}_o^{l,(t)})^2] + \gamma\frac{1}{|\mathcal{U}_o^{l,(t)}|}\mathrm{S}[(\mathcal{G}^l)^2 * \phi((U_o^{l,(t)})^2(\Sigma_o^{l,(t)})^2({V_o^{l,(t)}}^\top)^2)].
\end{aligned} \tag{9}$$

B. Algorithm A

We provide the heuristic compression algorithm in Algorithm. A.

Algorithm A Heuristic compression algorithm

Input: A single layer $\mathcal{W}^l$, average gradient of weight $\mathcal{G}^l$, target compression rate R_a^l, calculation interval T.

Output: The compressed layer $\overline{\mathcal{W}}^l$.

1: Initialize the set of compression unit index $\mathcal{U}^l$, whose corresponding unit number is $c^l + r^l$.
2: Initialize $\mathcal{W}^{l,0} = \mathcal{W}^l$, current compression rate $R^l = 0$, current step $t = 1$, removed input channels set $\mathcal{U}_{cp}^l = \emptyset$.
3: **while** true **do**
4:     **for** each compression unit index o in $\mathcal{U}^{l,(t)}$ **do**
5:         $P_o^{l,(t)} = I_o^{l,(t)} + \gamma \frac{1}{|\mathcal{U}^{l,(t)}|} \sum_{i \in \mathcal{U}^{l,(t)} \setminus o} I_{i|o}^{l,(t)}$.
6:     **end for**
7:     $\overline{\mathcal{W}}^{l,(t)} = \overline{\mathcal{W}}^{l,(t-1)}$.
8:     $\mathcal{U}^{l,(t+1)} = \mathcal{U}^{l,(t)}$.
9:     **for** T least important compression units index o in $\mathcal{U}^{l,(t)}$ **do**
10:         $\overline{\mathcal{W}}^{l,(t)} = f(\overline{\mathcal{W}}^{l,(t)}, o)$.
11:         $\mathcal{U}^{l,(t+1)} = \mathcal{U}^{l,(t+1)} \setminus o$.
12:         Compute R^l via Eq. 10 in this material.
13:         **if** o belongs to channel pruning **then**
14:             Add o to $\mathcal{U}_{cp}^l$.
15:         **end if**
16:         **if** $R^l >= R_a^l$ **then**
17:             **return** $\overline{\mathcal{W}}^{l,(t)}$.
18:         **end if**
19:         **if** $\frac{|\mathcal{U}_{cp}^l|}{c^l} >= R_a^l$ **then**
20:             $\overline{\mathcal{W}}^l = \mathcal{W}^l$.
21:             **for** each compression unit index o in $\mathcal{U}_{cp}^l$ **do**
22:                 $\overline{\mathcal{W}}^l = f(\overline{\mathcal{W}}^l, o)$.
23:             **end for**
24:             **return** $\overline{\mathcal{W}}^l$
25:         **end if**
26:     **end for**
27:     $t = t + 1$.
28: **end while**

Lines 19-25 in the above algorithm are based on the following analysis. According to the definition of the compression rate:

$$R^l = \begin{cases} 1 - \dfrac{(r^l - t_2) * [(c^l - t_1) * k^l * k^l + n^l]}{n^l * c^l * k^l * k^l}, & t_2 \neq 0; \\ \dfrac{t_1}{c^l}, & t_2 = 0, \end{cases} \tag{10}$$

if we remove a less number of singular values (t_2 is smaller but not equal to zero), the SVD-decomposition will increase the number of parameters, which perhaps leads to extra channel pruning (t_1 is larger) to achieve target compression rate. In contrast, if we only consider channel pruning (*i.e.,* $t_2 = 0$), t_1 will be smaller than the above situation, which keeps more information to achieve the target compression rate. Therefore, during the compression process, if the weight only compressed by removing input channels has reached the target compression rate (*i.e.,* $\frac{t_1}{c^l}$ larger than the target compression rate), we will only adopt the channel pruning to compress it.

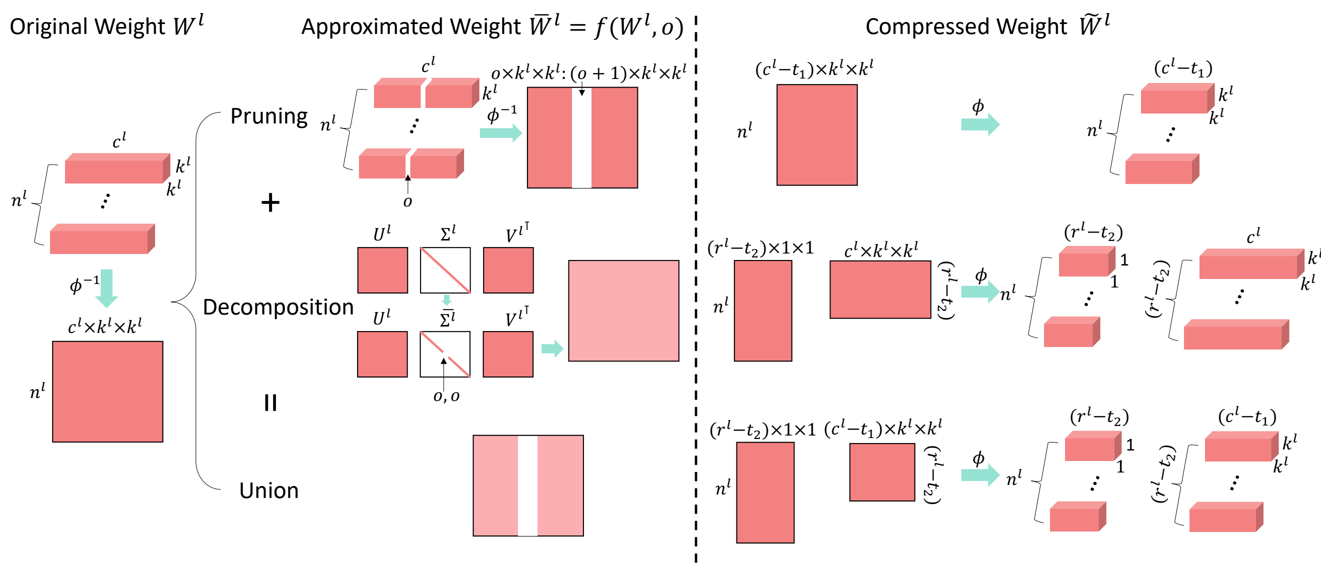

Figure A: Visualization of the compression process.

C. Visualization of the Compression Process

During the compression process, the approximated weight $\overline{\mathcal{W}}^l$ is used to compute the importance metric. After finishing the compression, we transform the approximated weight $\overline{\mathcal{W}}^l$ to compressed weight $\widetilde{\mathcal{W}}^l$ to the initial weight for the compressed network and then fine-tune the network. This process is demonstrated in the following figure.

D. More Comparison with State-of-the-Art Methods

We compare our method with other methods based on single compression operations for VGG-16 and ResNet-50. As shown in the following Tab. A, compared to GDP [4], our method achieves better performance (69.73% vs. 67.51%) with higher FLOPs reduction (77.5% vs. 75.5%). Meanwhile, compared to [4, 6, 7, 1, 2, 3, 5], we also achieve better performance for ResNet-50, which is shown in the following Fig. B.

Model	Method	FLOPs (PR)	#Param. (PR)	Top-1 Acc%	Top-5 Acc%
VGG-16	Baseline	15.48B	138M	71.59	90.38
	ThiNet[6]	9.58B(38.1%)	131M(5.1%)	69.80	89.53
	CC($C_t = 0.5$)	**7.56B(52.4%)**	**131M(5.1%)**	**72.05**	**90.61**
	GDP[4]	7.5B(54.5%)	-	69.88	89.16
	GDP[4]	3.8B(75.5%)	-	67.51	87.95
	CC($C_t = 0.75$)	**3.48B(77.5%)**	**127M(8.0%)**	**69.73**	**89.39**

Table A: Comparison with single compression operations-based methods for VGG-16 on ImageNet2012.

We evaluate the generalization ability of our method on PASCAL VOC object detection task. We compress Faster-RCNN with ResNet-50 backbone on Pascal VOC and only obtain 0.85 mAP drop with 50% compression rate, which demonstrates that our method has a strong generalization ability for the detection task.

References

[1] Xiaohan Ding, Guiguang Ding, Yuchen Guo, and Jungong Han. Centripetal sgd for pruning very deep convolutional networks with complicated structure. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[2] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

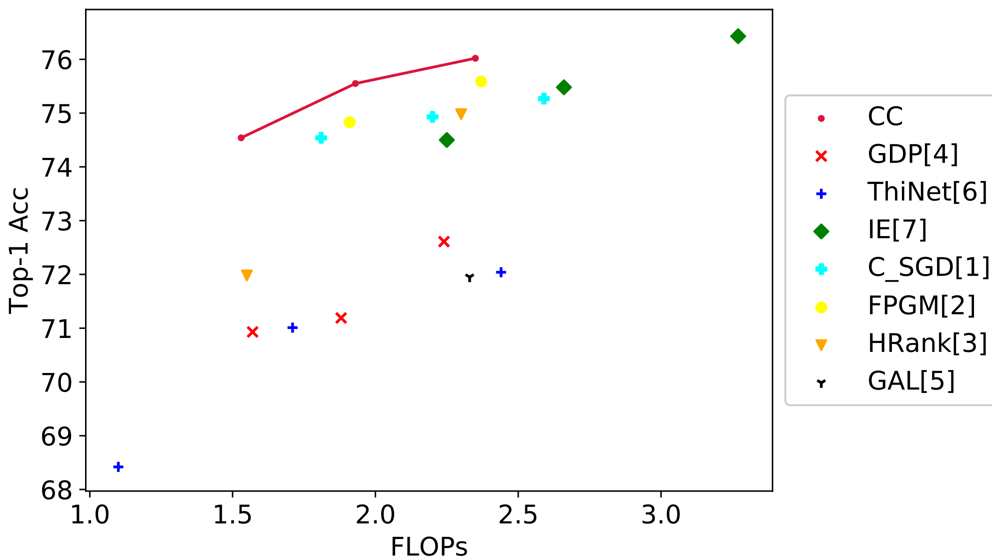

Figure B: Comparison with single compression operations-based methods for ResNet-50 on ImageNet2012.

[3] Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, and Ling Shao. Hrank: Filter pruning using high-rank feature map. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[4] Shaohui Lin, Rongrong Ji, Yuchao Li, Yongjian Wu, Feiyue Huang, and Baochang Zhang. Accelerating convolutional networks via global & dynamic filter pruning. *IJCAI*, pages 2425–2432, 2018. 4

[5] Shaohui Lin, Rongrong Ji, Chenqian Yan, Baochang Zhang, Liujuan Cao, Qixiang Ye, Feiyue Huang, and David Doermann. Towards optimal structured cnn pruning via generative adversarial learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[6] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. *International Conference on Computer Vision (ICCV)*, 2017. 4

[7] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4