Supplementary Materials for Towards Compact CNNs via Collaborative Compression

Yuchao Li\(^1\), 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:

\begin{equation}
    P_o^{l(t)} = P_o^{l(t)} + \gamma \frac{1}{|U^{l(t)}| - 1} \sum_{i \in U^{l(t)} \setminus o} I_{i|o}^{l(t)}.
\end{equation}

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

\begin{equation}
    \gamma \frac{1}{|U^{l(t)}| - 1} \sum_{i \in U^{l(t)} \setminus o} I_{i|o}^{l(t)} = \gamma \frac{1}{|U^{l(t)}|} \sum_{i \in U^{l(t)}} S[(G^l \ast (\tilde{W}_{i|o} - W^l)^2] = \frac{1}{|U^{l(t)}|} \sum_{i \in U^{l(t)}} S[(G^l \ast (W_{i|o}^l - W_{i|o}^l + W_{i|o}^l - W^l))^2]. \nonumber
\end{equation}

\begin{equation}
= \frac{1}{|U^{l(t)}|} \sum_{i \in U^{l(t)}} S[(G^l \ast (\theta_{i|o}^{l(t)} + \theta_{o}^{l(t)}))^2] = \gamma \frac{1}{|U^{l(t)}|} \sum_{i \in U^{l(t)}} S[(G^l \ast \theta_{i|o}^{l(t)})^2 + 2G^l \ast \theta_{i|o}^{l(t)} \ast \theta_{o}^{l(t)} + G^l \ast \theta_{o}^{l(t)} + (G^l \ast \theta_{o}^{l(t)})^2] = \gamma \frac{1}{|U^{l(t)}|} S(G^l)^2 \sum_{i \in U^{l(t)}} (\theta_{i|o}^{l(t)})^2 + \gamma S(G^l)^2 \ast \theta_{o}^{l(t)} \ast \sum_{i \in U^{l(t)}} \theta_{i|o}^{l(t)} + \gamma S(G^l \ast \theta_{o}^{l(t)})^2, \nonumber
\end{equation}

where \( \theta_{i|o}^{l(t)} = \tilde{W}_{i|o}^{l(t)} - W_{o}^{l(t)}, \theta_{o}^{l(t)} = W_{o}^{l(t)} - W^l \). \ast represents element-wise multiplication, and \( U^{l(t)} = U^{l(t)} \setminus o \). Then, we consider the compression units in channel pruning \( U_{\text{cp}|o}^{l(t)} \) and tensor decomposition \( U_{\text{td}|o}^{l(t)} \) separately, where \( U_{\text{cp}|o}^{l(t)} \cup U_{\text{td}|o}^{l(t)} = U_{o}^{l(t)} \). 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.
Therefore, the second item in Eq. 1 for channel pruning becomes:

\[
\frac{1}{K_{l}^{(t)}} \sum_{i \epsilon \mathcal{U}_{lp}^{(t)}} l_{i}^{(t)} = \frac{1}{K_{l}^{(t)}} \sum_{i \epsilon \mathcal{U}_{lp}^{(t)}} \theta_{i}^{(t)} \sum_{i \epsilon \mathcal{U}_{lp}^{(t)}} \phi(-U_{i}^{(t)} \Sigma_{i}^{(t)} V_{i}^{(t)^{T}})
\]

where \( \Sigma_{o}^{(t)} = \phi(U_{o}^{(t)} \Sigma_{o}^{(t)} V_{o}^{(t)^{T}}) \) is the SVD of \( \phi^{-1}(W_{o}^{(t)}) \). The second item of Eq. 2 can also be rewritten as:

\[
\frac{2}{K_{l}^{(t)}} \sum_{i \epsilon \mathcal{U}_{l}^{(t)}} \theta_{i}^{(t)} \sum_{i \epsilon \mathcal{U}_{l}^{(t)}} \phi(-U_{i}^{(t)} \Sigma_{i}^{(t)} V_{i}^{(t)^{T}})
\]

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

\[
\frac{1}{K_{l}^{(t)}} \sum_{i \epsilon \mathcal{U}_{l}^{(t)}} l_{i}^{(t)} = \frac{1}{K_{l}^{(t)}} \sum_{i \epsilon \mathcal{U}_{l}^{(t)}} \theta_{i}^{(t)} \sum_{i \epsilon \mathcal{U}_{l}^{(t)}} \phi(-U_{i}^{(t)} \Sigma_{i}^{(t)} V_{i}^{(t)^{T}})
\]

Finally, the importance metric can be formulated as follows:

\[
P_{o}^{(t)} = l_{o}^{(t)} + \frac{1}{K_{l}^{(t)}} \sum_{i \epsilon \mathcal{U}_{l}^{(t)}} l_{i}^{(t)} + \gamma \frac{1}{K_{l}^{(t)}} \sum_{i \epsilon \mathcal{U}_{l}^{(t)}} l_{i}^{(t)}
\]

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

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

(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., $\frac{1}{2}$ larger than the target compression rate), we will only adopt the channel pruning to compress it.
C. Visualization of the Compression Process

During the compression process, the approximated weight $\hat{W}^i$ is used to compute the importance metric. After finishing the compression, we transform the approximated weight $\hat{W}^i$ to compressed weight $\tilde{W}^i$ 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.

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

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


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


