Supplementary Material: Redistribution of Weights and Activations for AdderNet Quantization

Ying Nie 1  Kai Han 1,2  Haikang Diao 3  Chuanjian Liu 1  Enhua Wu 2,4  Yunhe Wang 1

1 Huawei Noah’s Ark Lab  
2 State Key Lab of Computer Science, ISCAS & UCAS  
3 School of Integrated Circuits, Peking University  
4 University of Macau  
{ying.nie, kai.han, yunhe.wang}@huawei.com

A Appendix

A.1 Proof of Proposition 2

Proof. For a weight \( W \in \mathbb{R}^{d \times d \times c_{in} \times c_{out}} \) and an input activation \( X \in \mathbb{R}^{h \times w \times c_{in}} \), the FLOPs of \( X \oplus W \) is \( 2hw(c_{in}d^2 + 1)c_{out} \). In addition, the FLOPs required to calculate \( \bar{w} \) and \( \bar{x} \) are \( c_{in}d^2c_{out} \) and \( ghwc_{in} \), respectively. Besides, the size of the output activation can be calculated by:

\[
h_{out} = \lfloor (h - d + 2 \times \text{padding})/\text{stride} + 1 \rfloor, \quad w_{out} = \lfloor (w - d + 2 \times \text{padding})/\text{stride} + 1 \rfloor. \tag{1}
\]

Without loss of generality, assume that the values of padding and stride are both 1, therefore, the FLOPs required to de-quantize the output activation is \( h_{out}w_{out}c_{out} = (h - d + 3)(w - d + 3)c_{out} \). Finally, all FLOPs required for a layer quantization is:

\[
FLOPs_{all} = 2hw(c_{in}d^2 + 1)c_{out} + c_{in}d^2c_{out} + ghwc_{in} + (h - d + 3)(w - d + 3)c_{out}. \tag{2}
\]

Without loss of generality, assume that \( d = 3, c_{in} = c_{out} = c, \) and \( h = w = k \), then

\[
FLOPs_{all} = 18k^2c^2 + gk^2c + 9c^2 + 3k^2c. \tag{3}
\]

Compared with only one scale (\( g = 1 \)), FLOPs are increased by \( r \) when adopting multiple scales (\( g \geq 2 \)):

\[
r = \frac{(g - 1)k^2c}{18k^2c^2 + k^2c + 9c^2 + 3k^2c} = \frac{(g - 1)k^2}{(18k^2 + 9)c + 4k^2} \approx \frac{g - 1}{18c + 4}, \tag{4}
\]

As we discussed in the section of experiments, the value of \( g \) that we adopt is 4. In this case, Eq. 4 can be further simplified to

\[
r \approx \frac{g - 1}{18c + 4} = \frac{3}{18c + 4} \approx \frac{1}{6c + 1}. \tag{5}
\]

Considering that the magnitude of \( c \) is generally in the tens or hundreds of common neural networks, thus the value of \( r \) is very small. Therefore, the increase in FLOPs brought by the scheme of group-shared scales is negligible.

A.2 Full-precision Results

In the section of experiments, we re-trained multiple full-precision adder networks on various datasets. The full-precision results on CIFAR-10 and CIFAR-100 are reported in Table 1 and the full-precision results on ImageNet are reported in Table 2 both denoted by AddNN. The results of AddNN are basically consistent with the results in [1]. The baseline results of convolutional neural network (CNN) and binary neural network (BNN) are cited from [1].

*Corresponding author

Table 1: Full-precision results on CIFAR-10 and CIFAR-100 datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th># Mul.</th>
<th># Add.</th>
<th># XNOR.</th>
<th>CIFAR-10 (%)</th>
<th>CIFAR-100 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-Small</td>
<td>CNN</td>
<td>0.65G</td>
<td>0.65G</td>
<td>0</td>
<td>93.80</td>
<td>72.73</td>
</tr>
<tr>
<td></td>
<td>BNN</td>
<td>0.05G</td>
<td>0.65G</td>
<td>0.60G</td>
<td>89.80</td>
<td>67.24</td>
</tr>
<tr>
<td></td>
<td>AddNN</td>
<td>0.05G</td>
<td>1.25G</td>
<td>0</td>
<td>93.44</td>
<td>73.60</td>
</tr>
<tr>
<td>ResNet-20</td>
<td>CNN</td>
<td>41.17M</td>
<td>41.17M</td>
<td>0</td>
<td>92.25</td>
<td>68.14</td>
</tr>
<tr>
<td></td>
<td>BNN</td>
<td>0.45M</td>
<td>41.17M</td>
<td>40.72M</td>
<td>84.87</td>
<td>54.14</td>
</tr>
<tr>
<td></td>
<td>AddNN</td>
<td>0.45M</td>
<td>81.89M</td>
<td>0</td>
<td>91.42</td>
<td>67.59</td>
</tr>
<tr>
<td>ResNet-32</td>
<td>CNN</td>
<td>69.12M</td>
<td>69.12M</td>
<td>0</td>
<td>93.29</td>
<td>69.74</td>
</tr>
<tr>
<td></td>
<td>BNN</td>
<td>0.45M</td>
<td>69.12M</td>
<td>68.67M</td>
<td>86.74</td>
<td>56.21</td>
</tr>
<tr>
<td></td>
<td>AddNN</td>
<td>0.45M</td>
<td>137.79M</td>
<td>0</td>
<td>92.72</td>
<td>70.17</td>
</tr>
</tbody>
</table>

Table 2: Full-precision results on ImageNet.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th># Mul.</th>
<th># Add.</th>
<th># XNOR.</th>
<th>Top-1 Acc (%)</th>
<th>Top-5 Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-18</td>
<td>CNN</td>
<td>1.8G</td>
<td>1.8G</td>
<td>0</td>
<td>69.8</td>
<td>89.1</td>
</tr>
<tr>
<td></td>
<td>BNN</td>
<td>0.1G</td>
<td>1.8G</td>
<td>1.7G</td>
<td>51.2</td>
<td>73.2</td>
</tr>
<tr>
<td></td>
<td>AddNN</td>
<td>0.1G</td>
<td>3.5G</td>
<td>0</td>
<td>67.9</td>
<td>87.8</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>CNN</td>
<td>3.9G</td>
<td>3.9G</td>
<td>0</td>
<td>76.2</td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>BNN</td>
<td>0.1G</td>
<td>3.9G</td>
<td>3.8G</td>
<td>55.8</td>
<td>78.4</td>
</tr>
<tr>
<td></td>
<td>AddNN</td>
<td>0.1G</td>
<td>7.6G</td>
<td>0</td>
<td>75.0</td>
<td>91.9</td>
</tr>
</tbody>
</table>

A.3 Analysis on the Ratio of Discarded Outliers

As we discussed in the subsection of outliers clamp for activations, the value \( r_x = \tilde{X}[[\alpha * (n - 1)]] \) is selected as the range of activations for the calculation of scale, where \( \alpha \in (0, 1] \) is a hyper-parameter controlling the ratio of discarded outliers in activations. We supplement the ablation study of this ratio with 4-bit quantized adder ResNet-20 network on CIFAR-100 dataset.

Table 3: Analysis on the ratio of discarded outliers in activations.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>0.9985</th>
<th>0.9990</th>
<th>0.9995</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc (%)</td>
<td>67.29</td>
<td>67.35</td>
<td>67.11</td>
<td>65.17</td>
</tr>
</tbody>
</table>

As shown in Table 3, \( \alpha = 1 \) means that the scheme of outliers clamp for activations is not adopted, resulting in a significantly degraded quantized accuracy. The quantized accuracy can be improved with an appropriate \( \alpha \).

A.4 Quantization Results on Adder Vision Transformers

We also try the proposed quantization method on adder vision transformers. We re-train the full-precision adder DeiT-T for 400 epochs from scratch on ImageNet dataset following, and the final top-1 accuracy of the full-precision adder DeiT-T is 68.3%. For the next quantization step, the number of groups we use is 4, the hyper-parameter \( \alpha \) controlling the ratio of discarded outliers in activations is set to 0.9992. The accuracy drops after post-training quantization are reported in Table 4. The advantage of the proposed method over QSSF is significant. For example, at the case of W4A4, the accuracy drop of our method is 8.7%, which is much lower than the 16.3% of QSSF.

A.5 Quantization Results on Lower-bit

We supplement the 3-bit PTQ quantization experiment of adder ResNet-20 on CIFAR-100 dataset. Besides, the comparisons with more CNN quantization methods are also supplemented. The detailed accuracy drops are reported in Table.

A.6 Distribution of the Weights and Activations

In Figure, we visualize the histogram of the weights and activations in AdderNet. The input full-precision (FP) activations and weights in pre-trained AdderNet show a significant difference,
which pose a huge challenge for AdderNet quantization. Other AdderNet quantization methods \( [6, 2] \) fail to deal with this challenge, leading to the phenomenon of over clamp and bits waste, further resulting in a poor quantized accuracy. In contrast, our quantization method can effectively address this challenge by the redistribution of full-precision weights and activations, resulting in a good quantized accuracy. One-shared scale is adopted here for the simplification of visualization, and symmetric 4-bit quantization is taken as an example.

### A.7 Limitations and Societal Impacts

Our AdderNet quantization method has one major limitation: as the number of bits decreases, the accuracy loss of the quantization model will increase. Therefore, quantization-aware training is necessary for the low bits, which is time consuming and computationally consuming.

As for the societal impacts, the proposed quantization method can further reduce the energy consumption of AdderNet with a lower quantized accuracy loss. The low power devices equipped with quantized AdderNet can be deployed to surveillance scenario. If used improperly, there may be a risk of information leakage.

### References


Figure 1: Distribution of the weights and activations in AdderNet.