Supplementary Materials for MST-compression: Compressing and Accelerating Binary Neural Networks

A. Proof of Eq. (2)

According to the convolution definition, the output of the channel i as the following

$$Y_i = \left(\sum_{j=1}^{C_{in}MM} \mathcal{A}_{ij}^b * \mathcal{W}_{ij}^b \right) \odot \alpha \tag{8}$$

There are $C_{in} \times M \times M$ multiplications, and these multiplications' output is -1 or 1. Assuming that there are A multiplications with the output -1 and B multiplication with output 1, we have $A + B = C_{in} \times M \times M$. Thus, Y_i can be derived as

$$Y_i = A - B = 2A - C_{in} \times M \times M. \tag{A1}$$

In addition, because $\mathcal{A}_{ij}^b$ and $\mathcal{W}_{ij}^b$ are binarized, A can be calculated as

$$A = \sum_{j=1}^{C_{in}MM} \text{XNOR}(\mathcal{A}_{ij}^b, \mathcal{W}_{ij}^b). \tag{A2}$$

Finally, we have Eq. (2) by replacing A with Eq. (A2) as

$$Y_i = (2 \sum_{j=1}^{C_{in}MM} \text{XNOR}(\mathcal{A}_{ij}^b, \mathcal{W}_{ij}^b) - C_{in} \times M \times M) \odot \alpha. \tag{2}$$

B. Proof of Eq. (3)

Assuming that $\mathcal{S}$ includes weight values of the channel j, which are similar to the weights of the channel i (compared one-one respectively). $\mathcal{D}$ includes weight values of the channels j, which are different from the weights of the channel i (compared one-one respectively), $|\mathcal{D}| = d_{ij}$. $\mathcal{A}_s^b$ and $\mathcal{A}_d^b$ are input activations for $\mathcal{S}$ and $\mathcal{D}$, respectively. P_j can be as

$$P_j = \sum_{\mathcal{W}_s \in \mathcal{S}} \text{XNOR}(\mathcal{A}_s^b, \mathcal{W}_s) + \sum_{\mathcal{W}_d \in \mathcal{D}} \text{XNOR}(\mathcal{A}_d^b, \mathcal{W}_d). \tag{A3}$$

Because input activation of the channel i is the same as that of the channel j. Suppose $\bar{\mathcal{D}}$ includes weights of the channel i, which are different from that of the channel j, $|\mathcal{D}| = |\bar{\mathcal{D}}| = d_{ij}$. We can have P_i as

$$P_i = \sum_{\mathcal{W}_s \in \mathcal{S}} \text{XNOR}(\mathcal{A}_s^b, \mathcal{W}_s) + \sum_{\bar{\mathcal{W}}_d \in \bar{\mathcal{D}}} \text{XNOR}(\mathcal{A}_d^b, \bar{\mathcal{W}}_d), \tag{A4}$$

In consequence, $\sum_{\mathcal{W}_s \in \mathcal{S}} \text{XNOR}(\mathcal{A}_s^b, \mathcal{W}_s)$ can be calculated as,

$$\sum_{\mathcal{W}_s \in \mathcal{S}} \text{XNOR}(\mathcal{A}_s^b, \mathcal{W}_s) = P_i - \sum_{\bar{\mathcal{W}}_d \in \bar{\mathcal{D}}} \text{XNOR}(\mathcal{A}_d^b, \bar{\mathcal{W}}_d), \tag{A5}$$

and $\forall$ input activations, based the characteristics of XNOR operation, we have

$$\sum_{\mathcal{W}_d \in \mathcal{D}} \text{XNOR}(\mathcal{A}_d^b, \mathcal{W}_d) + \sum_{\bar{\mathcal{W}}_d \in \bar{\mathcal{D}}} \text{XNOR}(\mathcal{A}_d^b, \bar{\mathcal{W}}_d) = d_{ij}. \tag{A6}$$

Use Eq. (A5) and Eq. (A6), we can reformulate the Eq. (A3) as

$$P_j = P_i - d_{ij} + 2 \sum_{\mathcal{W}_d \in \mathcal{D}} \text{XNOR}(\mathcal{A}_d^b, \mathcal{W}_d). \tag{A7}$$

In Sec. 2, we have $P_{ij} = \sum_{\mathcal{W}_d \in \mathcal{D}} \text{XNOR}(\mathcal{A}_d^b, \mathcal{W}_d)$. Thus, we finally have the following equation.

$$Y_j = 2(P_i - d_{ij} + 2P_{ij}) - C_{in} \times M \times M. \tag{3}$$

C. Additional results

Effect of the number of centers. In this section, we provide an additional experimental results related to the effect of the number of initial centers for the training. In particular, we do the training on VGG-small model and CIFAR-10 dataset with different number of centers, while λ is fixed at 4e-6. Besides, each number of centers, we execute the training three times and get the mean value for the report.

Table A1 provides the MST depth, number of parameters, bit-ops and accuracy *w.r.t.* different number of centers. Accordingly, the MST depth, number of parameters and bit-ops tend to increase as the number of centers increases.

#centers	MST-depth	#Params (Mbit)	#Bit-Ops (GOps)	Top-1 Acc. mean ± std (%)
1	22.3	0.545	0.119	91.49 ± 0.04
2	30.3	0.550	0.118	91.45 ± 0.08
4	47.7	0.574	0.125	91.42 ± 0.06
6	60.3	0.581	0.130	91.53 ± 0.07
8	73.0	0.607	0.136	91.49 ± 0.04

Table A1. Accuracy, MST depth, number of parameters and bit-Ops *w.r.t.* different number of centers on CIFAR-10 VGG-small model.

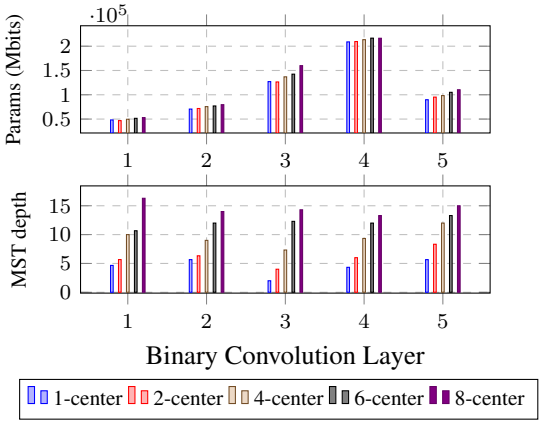

Figure A1. Number of parameters and MST depth on each convolution layer *w.r.t.* different number of centers.

Specifically, when the number of centers changes from 1 to 8, the MST depth increases $3.27\times$, the number of parameters and bit-ops increase $1.11\times$ and $1.14\times$, respectively. Meanwhile, accuracy barely changes with different number of centers. For each binary convolution layer, as shown in Figure A1, as the number of centers increases, both the MST depth and number of parameters also increase. These findings suggest that opting for a single center is the most effective strategy to minimize MST depth, parameters, and bit-ops while preserving accuracy.