

## A Appendix / supplemental material

### A.1 Proof [Proposition 2]

Consider a one layer network without bias term. The input dimension of it is  $n$ . The output dimension of it is  $m$ . We can know the shape of the weight matrix  $W$  is  $\mathbb{R}^{m \times n}$ . Since we have  $N$  data points, the input matrix  $X \in \mathbb{R}^{N \times n}$ .

#### Forward FLOPs with full matrix

*Proof.* When using full weight matrix, the first step is to compute the output  $O \in \mathbb{R}^{N \times m}$  as

$$O = XW^T. \quad (8)$$

The FLOPs of this step is  $Nm(2n - 1)$ . Then we calculate the loss as

$$\mathcal{J} = \|O - Y\|_F^2, \quad (9)$$

where  $Y \in \mathbb{R}^{N \times m}$  is the label matrix. The FLOPs for this step is  $3Nm - 1$ . Therefore, the FLOPs of the forward computation is

$$Nm(2n - 1) + (3Nm - 1) = \mathcal{O}(2Nm(n + 1)). \quad (10)$$

□

#### Backward FLOPs with full matrix

*Proof.* In the backward process, we need to calculate the gradient of  $\mathcal{J}$  on  $W$ . Using chain rule, the first step is to compute

$$\frac{\partial \mathcal{J}}{\partial O} = 2(O - Y). \quad (11)$$

Since  $O - Y$  has been calculated in the forward pass, the FLOPs is  $Nm$ . The gradient of  $W$  is

$$\frac{\partial \mathcal{J}}{\partial W} = \left( \frac{\partial \mathcal{J}}{\partial O} \right)^T X. \quad (12)$$

The FLOPs for this step is  $mn(2N - 1)$ . Therefore, the FLOPs for the backward pass is

$$Nm + mn(2N - 1) = \mathcal{O}(Nm(2n + 1)). \quad (13)$$

□

**Forward FLOPs with sparse matrix** With Kronecker product decomposition, we replace  $W$  by  $\sum_{i=1}^r (S \odot A_i) \otimes B_i$ .  $S$  and  $A_i \in \mathbb{R}^{m_1 \times n_1}$ ,  $B_i \in \mathbb{R}^{m_2 \times n_2}$ , where  $m_1 m_2 = m$ ,  $n_1 n_2 = n$ .

*Proof.* In the forward pass, we need to firstly reshape  $X \in \mathbb{R}^{N \times n}$  into  $\text{reshape}(X) \in \mathbb{R}^{n_2 \times N n_1}$ . Then we calculate  $B_i \text{reshape}(X) \in \mathbb{R}^{m_2 \times N n_1}$  with FLOPs  $N n_1 m_2 (2n_2 - 1)$ . The result is reshape into  $\text{reshape}(B_i \text{reshape}(X)) \in \mathbb{R}^{N m_2 \times n_1}$ .

Then we calculate  $S \odot A_i \in \mathbb{R}^{n_1 \times m_1}$  with FLOPs  $m_1 n_1$ . After this, we get

$$O_i = \text{reshape}(B_i \text{reshape}(X))(S \odot A_i)^T \in \mathbb{R}^{N m_2 \times m_1}, \quad (14)$$

with FLOPs  $N m_1 m_2 (2n_1 - 1)$ . We denote  $O$  as the output of the layer, which is to say

$$O = \sum_{i=1}^r O_i = \sum_{i=1}^r \text{reshape}(B_i \text{reshape}(X))(S \odot A_i)^T. \quad (15)$$

The total FLOPs to get  $O$  is

$$r(N m_1 m_2 (2n_1 - 1) + m_1 n_1 + N n_1 m_2 (2n_2 - 1)) + (r - 1)Nm \quad (16)$$

Then we reshape  $O$  in to  $\text{reshape}(O) \in \mathbb{R}^{N \times m}$ . The loss is calculated as

$$\mathcal{J} = \|\text{reshape}(O) - Y\|_F^2 \quad (17)$$

The FLOPs for this step is  $3Nm - 1$ . Therefore the FLOPs of the forward computation is

$$\begin{aligned} & r(2N m_1 m_2 n_1 - N m_1 m_2 + m_1 n_1 + 2N m_1 n_1 n_2 - N m_2 n_1) + (r - 1)Nm + 3Nm - 1 \\ & = \mathcal{O}(2N r m_1 n_1 (m_2 + n_2) - N r (m + 2m_2 n_1) + 3Nm) \end{aligned} \quad (18)$$

□

### Backward FLOPs with sparse matrix

*Proof.* In the backward process, we need to calculate the gradient of  $\mathcal{J}$  on  $S, A_i$  and  $B_i$ . Using chain rule, the first step is to compute

$$\frac{\partial \mathcal{J}}{\partial \text{reshape}(O)} = 2(\text{reshape}(O) - Y). \quad (19)$$

Since  $\text{reshape}(O) - Y$  has been calculated in the forward pass, the FLOPs is  $Nm$ . Then we reshape it into  $\frac{\partial \mathcal{J}}{\partial O} \in \mathbb{R}^{Nm_2 \times m_1}$ . Then we can get

$$\frac{\partial \mathcal{J}}{\partial (S \odot A_i)} = \left( \frac{\partial \mathcal{J}}{\partial O} \right)^T \text{reshape}(B_i \text{reshape}(X)). \quad (20)$$

Since  $\text{reshape}(B_i \text{reshape}(X))$  has been obtained in the forward pass, the FLOPs for this step is  $m_1 n_1 (2Nm_2 - 1)$ . To get the gradient on  $S$  and  $A_i$ , we have

$$\frac{\partial \mathcal{J}}{\partial S} = \sum_{i=1}^r \frac{\partial \mathcal{J}}{\partial (S \odot A_i)} \odot A_i, \quad (21)$$

with FLOPs  $rm_1 n_1 + (r - 1)m_1 n_1$ , and

$$\frac{\partial \mathcal{J}}{\partial A_i} = \frac{\partial \mathcal{J}}{\partial (S \odot A_i)} \odot S, \quad (22)$$

with FLOPs  $m_1 n_1$ . The gradient on  $\text{reshape}(B_i \text{reshape}(X))$  is

$$\frac{\partial \mathcal{J}}{\partial \text{reshape}(B_i \text{reshape}(X))} = \frac{\partial \mathcal{J}}{\partial O} (S \odot A_i). \quad (23)$$

The FLOPs for this step is  $Nm_2 n_1 (2m_1 - 1)$ . We reshape the gradient into  $\frac{\partial \mathcal{J}}{\partial B_i \text{reshape}(X)} \in \mathbb{R}^{m_2 \times Nn_1}$ . So, we can get the gradient on  $B_i$  as

$$\frac{\partial \mathcal{J}}{\partial B_i} = \frac{\partial \mathcal{J}}{\partial B_i \text{reshape}(X)} \text{reshape}(X)^T, \quad (24)$$

with FLOPs of  $m_2 n_2 (2Nn_1 - 1)$ . Therefore, we can get the total FLOPs for the backward pass as

$$\begin{aligned} & Nm + rm_1 n_1 (2Nm_2 - 1) + rm_1 n_1 + (r - 1)m_1 n_1 + rm_1 n_1 + rNm_2 n_1 (2m_1 - 1) + rm_2 n_2 (2Nn_1 - 1) \\ & = \mathcal{O}(Nm + Nr(4m_1 m_2 n_1 - m_2 n_1 + 2m_2 n_1 n_2)) \end{aligned} \quad (25)$$

□

### A.2 Proof [Proposition 3]

Consider a two-layer network without bias term. The input dimension of the first linear layer is  $m^{[1]}$ . The output dimension of the linear layer is  $m^{[2]}$ . So, the input dimension for the second linear layer is  $m^{[2]}$ . The output dimension of the linear layer is  $m^{[3]}$ . With the network architecture, we can know that  $W^{[1]} \in \mathbb{R}^{m^{[2]} \times m^{[1]}}$  and  $W^{[2]} \in \mathbb{R}^{m^{[3]} \times m^{[2]}}$ . Since we have  $N$  data points, for the input matrix of the first layer denoted by  $X^{[1]}$ , we have,  $X^{[1]} \in \mathbb{R}^{N \times m^{[1]}}$ . We use  $\sigma$  to represent the activation function for the first layer.

#### Forward FLOPs with full matrix

*Proof.* When using full weight matrix, the first step is to compute the pre-activated output  $O^{[1]} \in \mathbb{R}^{N \times m^{[2]}}$  as

$$O^{[1]} = X^{[1]} W^{[1]T}. \quad (26)$$

The FLOPs of this step is  $Nm^{[2]}(2m^{[1]} - 1)$ . Then, we compute the activation  $X^{[2]} \in \mathbb{R}^{N \times m^{[2]}}$  as follows,

$$X^{[2]} = \sigma(O^{[1]}). \quad (27)$$

The FLOPs for calculating the activation is  $Nm^{[2]}$ . For the second layer, we need to compute  $O^{[2]} \in \mathbb{R}^{N \times m^{[3]}}$  as  $O^{[2]} = X^{[2]} W^{[2]T}$ . The FLOPs are  $Nm^{[3]}(2m^{[2]} - 1)$ . The last step is calculating the loss as

$$\mathcal{J} = \left\| O^{[2]} - Y \right\|_F^2, \quad (28)$$

where  $Y \in \mathbb{R}^{N \times m^{[3]}}$  is the label matrix. The FLOPs for this step is  $3Nm^{[3]} - 1$ . Therefore, the FLOPs of the forward computation is

$$\begin{aligned} & Nm^{[2]}(2m^{[1]} - 1) + Nm^{[2]} + Nm^{[3]}(2m^{[2]} - 1) + 3Nm^{[3]} - 1 = 2Nm^{[1]}m^{[2]} + 2Nm^{[2]}m^{[3]} + 2Nm^{[3]} - 1 \\ & = \mathcal{O}\left(2N(m^{[1]}m^{[2]} + m^{[2]}m^{[3]} + m^{[3]})\right). \end{aligned} \quad (29)$$

□

### Backward FLOPs with full matrix

*Proof.* In the backward process, we need to calculate the gradient of  $\mathcal{J}$  on  $W^{[1]}$  and  $W^{[2]}$ . Using chain rule, the first step is to compute

$$\frac{\partial \mathcal{J}}{\partial O^{[2]}} = 2(O^{[2]} - Y). \quad (30)$$

Since  $O^{[2]} - Y$  has been calculated in the forward pass, the FLOPs is  $Nm^{[3]}$ . The gradient of  $W^{[2]}$  is

$$\frac{\partial \mathcal{J}}{\partial W^{[2]}} = \left( \frac{\partial \mathcal{J}}{\partial O^{[2]}} \right)^T X^{[2]}. \quad (31)$$

The FLOPs for this step is  $m^{[2]}m^{[3]}(2N - 1)$ . To compute the gradient on  $W^{[1]}$ , we need to first compute

$$\frac{\partial \mathcal{J}}{\partial X^{[2]}} = \left( \frac{\partial \mathcal{J}}{\partial O^{[2]}} \right) W^{[2]}, \quad (32)$$

with the FLOPs  $Nm^{[2]}(2m^{[3]} - 1)$ . Then compute

$$\frac{\partial \mathcal{J}}{\partial O^{[1]}} = \left( \frac{\partial \mathcal{J}}{\partial X^{[2]}} \right) \odot (\sigma(O^{[1]})), \quad (33)$$

with the FLOPs  $Nm^{[2]}$ . The final step is

$$\frac{\partial \mathcal{J}}{\partial W^{[1]}} = \left( \frac{\partial \mathcal{J}}{\partial O^{[1]}} \right)^T X^{[1]} \quad (34)$$

The FLOPs is  $m^{[1]}m^{[2]}(2N - 1)$ . Therefore, the FLOPs for the backward pass is

$$\begin{aligned} & Nm^{[3]} + m^{[2]}m^{[3]}(2N - 1) + Nm^{[2]}(2m^{[3]} - 1) + Nm^{[2]} + m^{[1]}m^{[2]}(2N - 1) \\ &= 2Nm^{[1]}m^{[2]} + 4Nm^{[2]}m^{[3]} + Nm^{[3]} - m^{[1]}m^{[2]} - m^{[2]}m^{[3]} \\ &= O\left(N(2m^{[1]}m^{[2]} + 4m^{[2]}m^{[3]} + m^{[3]})\right). \end{aligned} \quad (35)$$

□

**Forward FLOPs with sparse matrix** With the Kronecker product decomposition, we replace  $W^{[1]}$  by  $\sum_{i=1}^{r^{[1]}} (S^{[1]} \odot A_i^{[1]}) \otimes B_i^{[1]}$  and  $W^{[2]}$  by  $\sum_{i=1}^{r^{[2]}} (S^{[2]} \odot A_i^{[2]}) \otimes B_i^{[2]}$ . In the first layer,  $S^{[1]}$  and  $A_i^{[1]} \in \mathbb{R}^{m_1^{[2]} \times m_1^{[1]}}$ ,  $B_i^{[1]} \in \mathbb{R}^{m_2^{[2]} \times m_2^{[1]}}$ , where  $m_1^{[1]}m_2^{[1]} = m^{[1]}$ ,  $m_1^{[2]}m_2^{[2]} = m^{[2]}$ . In the second layer,  $S^{[2]}$  and  $A_i^{[2]} \in \mathbb{R}^{m_1^{[3]} \times m_1^{[2]}}$ ,  $B_i^{[2]} \in \mathbb{R}^{m_2^{[3]} \times m_2^{[2]}}$ , where  $m_1^{[3]}m_2^{[3]} = m^{[3]}$ .

*Proof.* In the forward pass, we need to firstly reshape  $X^{[1]} \in \mathbb{R}^{N \times m^{[1]}}$  into  $\text{reshape}(X^{[1]}) \in \mathbb{R}^{m_2^{[1]} \times Nm_1^{[1]}}$ . Then we calculate  $B_i^{[1]} \text{reshape}(X^{[1]}) \in \mathbb{R}^{m_2^{[2]} \times Nm_1^{[1]}}$  with the FLOPs  $Nm_1^{[1]}m_2^{[2]}(2m_2^{[1]} - 1)$ . The result is reshaped into  $\text{reshape}(B_i^{[1]} \text{reshape}(X^{[1]})) \in \mathbb{R}^{Nm_2^{[2]} \times m_1^{[1]}}$ .

Then we calculate  $S^{[1]} \odot A_i^{[1]} \in \mathbb{R}^{m_1^{[2]} \times m_1^{[1]}}$  with the FLOPs  $m_1^{[1]}m_1^{[2]}$ . After this, we get

$$O_i^{[1]} = \text{reshape}(B_i^{[1]} \text{reshape}(X^{[1]}))(S^{[1]} \odot A_i^{[1]})^T \in \mathbb{R}^{Nm_2^{[2]} \times m_1^{[2]}}, \quad (36)$$

with FLOPs  $Nm_2^{[2]}m_1^{[2]}(2m_1^{[1]} - 1)$ . We denote  $O^{[1]}$  as the pre-activated result of the first layer, which is to say

$$O^{[1]} = \sum_{i=1}^{r^{[1]}} O_i^{[1]} = \sum_{i=1}^{r^{[1]}} \text{reshape}(B_i^{[1]} \text{reshape}(X^{[1]}))(S^{[1]} \odot A_i^{[1]})^T. \quad (37)$$

The total FLOPs to get  $O^{[1]}$  is

$$\begin{aligned} & r^{[1]} \left( Nm_1^{[1]}m_2^{[2]}(2m_2^{[1]} - 1) + m_1^{[1]}m_1^{[2]} + Nm_2^{[2]}m_1^{[2]}(2m_1^{[1]} - 1) \right) + (r^{[1]} - 1)m^{[2]}N \\ &= r^{[1]} \left( 2Nm^{[1]}m_2^{[2]} + 2Nm^{[2]}m_1^{[1]} - Nm_2^{[2]}(m_1^{[1]} + m_1^{[2]}) + m_1^{[1]}m_1^{[2]} \right) + (r^{[1]} - 1)m^{[2]}N. \end{aligned} \quad (38)$$

The input for the second layer  $X^{[2]} \in \mathbb{R}^{Nm_2^{[2]} \times m_1^{[2]}}$  is obtained by

$$X^{[2]} = \sigma(O^{[1]}), \quad (39)$$

with FLOPs  $Nm_2^{[2]}m_1^{[2]} = Nm^{[2]}$ . Then we reshape it into  $\text{reshape}(X^{[2]}) \in \mathbb{R}^{m_2^{[2]} \times Nm_1^{[2]}}$ . We calculate  $B_i^{[2]} \text{reshape}(X^{[2]}) \in \mathbb{R}^{m_2^{[3]} \times Nm_1^{[2]}}$  with the FLOPs  $Nm_1^{[2]}m_2^{[3]}(2m_2^{[2]} - 1)$ . The result is reshaped into  $\text{reshape}(B_i^{[2]} \text{reshape}(X^{[2]})) \in \mathbb{R}^{Nm_2^{[3]} \times m_1^{[2]}}$ .

Similar to the first layer, we calculate  $S^{[2]} \odot A_i^{[2]} \in \mathbb{R}^{m_1^{[3]} \times m_1^{[2]}}$  with the FLOPs  $m_1^{[2]}m_1^{[3]}$ . After this, we get

$$O_i^{[2]} = \text{reshape}(B_i^{[2]} \text{reshape}(X^{[2]}))(S^{[2]} \odot A_i^{[2]})^T \in \mathbb{R}^{Nm_2^{[3]} \times m_1^{[3]}}, \quad (40)$$

with FLOPs  $Nm_2^{[3]}m_1^{[3]}(2m_1^{[2]} - 1)$ . We denote  $O^{[2]}$  as the output of the second layer, which is to say

$$O^{[2]} = \sum_{i=1}^{r^{[2]}} O_i^{[2]} = \sum_{i=1}^{r^{[2]}} \text{reshape}(B_i^{[2]} \text{reshape}(X^{[2]}))(S^{[2]} \odot A_i^{[2]})^T. \quad (41)$$

The total FLOPs to get  $O^{[2]}$  is

$$\begin{aligned} & r^{[2]} \left( Nm_1^{[2]}m_2^{[3]}(2m_2^{[2]} - 1) + m_1^{[2]}m_1^{[3]} + Nm_2^{[3]}m_1^{[3]}(2m_1^{[2]} - 1) \right) + (r^{[2]} - 1)m^{[3]}N \\ & = r^{[2]} \left( 2Nm^{[2]}m_2^{[3]} + 2Nm^{[3]}m_1^{[2]} - Nm_2^{[3]}(m_1^{[2]} + m_1^{[3]}) + m_1^{[2]}m_1^{[3]} \right) + (r^{[2]} - 1)m^{[3]}N. \end{aligned} \quad (42)$$

Then we shape  $O^{[2]}$  into  $\text{reshape}(O^{[2]}) \in \mathbb{R}^{N \times m^{[3]}}$ . The loss is calculated as

$$\mathcal{J} = \left\| \text{reshape}(O^{[2]}) - Y \right\|_F^2. \quad (43)$$

The FLOPs for this step is  $3Nm^{[3]} - 1$ . Therefore, the FLOPs of the forward computation is

$$\begin{aligned} & r^{[1]} \left( 2Nm^{[1]}m_2^{[2]} + 2Nm^{[2]}m_1^{[1]} - Nm_2^{[2]}(m_1^{[1]} + m_1^{[2]}) + m_1^{[1]}m_1^{[2]} \right) + (r^{[1]} - 1)m^{[2]}N + Nm^{[2]} + \\ & r^{[2]} \left( 2Nm^{[2]}m_2^{[3]} + 2Nm^{[3]}m_1^{[2]} - Nm_2^{[3]}(m_1^{[2]} + m_1^{[3]}) + m_1^{[2]}m_1^{[3]} \right) + (r^{[2]} - 1)m^{[3]}N + 3Nm^{[3]} - 1 \\ & = \mathcal{O} \left( r^{[1]} \left( 2Nm^{[1]}m_2^{[2]} + 2Nm^{[2]}m_1^{[1]} - Nm_2^{[2]}(m_1^{[1]} + m_1^{[2]}) \right) + r^{[2]} \left( 2Nm^{[2]}m_2^{[3]} + 2Nm^{[3]}m_1^{[2]} - Nm_2^{[3]}(m_1^{[2]} + m_1^{[3]}) \right) + Nm^{[2]} + 3Nm^{[3]} \right) \end{aligned} \quad (44)$$

Let  $C_1 = 2Nm^{[1]}m_2^{[2]} + 2Nm^{[2]}m_1^{[1]} - Nm_2^{[2]}(m_1^{[1]} + m_1^{[2]})$ ,  $C_2 = 2Nm^{[2]}m_2^{[3]} + 2Nm^{[3]}m_1^{[2]} - Nm_2^{[3]}(m_1^{[2]} + m_1^{[3]})$ , we have the FLOPs as  $\mathcal{O} \left( r^{[1]}C_1 + r^{[2]}C_2 + Nm^{[2]} + 3Nm^{[3]} \right)$ .  $\square$

### Backward FLOPs with sparse matrix

*Proof.* In the backward process, we need to calculate the gradient of  $\mathcal{J}$  on  $S^{[1]}$ ,  $A_i^{[1]}$ ,  $B_i^{[1]}$ ,  $S^{[2]}$ ,  $A_i^{[2]}$  and  $B_i^{[2]}$ . Using chain rule, the first step is to compute

$$\frac{\partial \mathcal{J}}{\partial \text{reshape}(O^{[2]})} = 2(\text{reshape}(O^{[2]}) - Y). \quad (45)$$

Since  $\text{reshape}(O^{[2]}) - Y$  has been calculated in the forward pass, the FLOPs is  $Nm^{[3]}$ . Then we reshape it into  $\frac{\partial \mathcal{J}}{\partial O^{[2]}} \in \mathbb{R}^{Nm_2^{[3]} \times m_1^{[3]}}$ . Then we can get

$$\frac{\partial \mathcal{J}}{\partial (S^{[2]} \odot A_i^{[2]})} = \left( \frac{\partial \mathcal{J}}{\partial O^{[2]}} \right)^T \text{reshape}(B_i^{[2]} \text{reshape}(X^{[2]})). \quad (46)$$

Since  $\text{reshape}(B_i^{[2]} \text{reshape}(X^{[2]}))$  has been obtained in the forward pass, the FLOPs for this step is  $m_1^{[2]}m_1^{[3]}(2Nm_2^{[3]} - 1)$ . To get the gradient on  $S^{[2]}$  and  $A_i^{[2]}$ , we have

$$\frac{\partial \mathcal{J}}{\partial S^{[2]}} = \sum_{i=1}^{r^{[2]}} \frac{\partial \mathcal{J}}{\partial (S^{[2]} \odot A_i^{[2]})} \odot A_i^{[2]}, \quad (47)$$

with FLOPs  $r^{[2]}m_1^{[2]}m_1^{[3]} + r^{[2]} - 1$ , and

$$\frac{\partial \mathcal{J}}{\partial A^{[2]}} = \frac{\partial \mathcal{J}}{\partial (S^{[2]} \odot A_i^{[2]})} \odot S^{[2]}, \quad (48)$$

with FLOPs  $m_1^{[2]}m_1^{[3]}$ . The gradient on  $\text{reshape}(B_i^{[2]}\text{reshape}(X^{[2]}))$  is

$$\frac{\partial \mathcal{J}}{\partial \text{reshape}(B_i^{[2]}\text{reshape}(X^{[2]}))} = \frac{\partial \mathcal{J}}{\partial O^{[2]}}(S^{[2]} \odot A_i^{[2]}). \quad (49)$$

The FLOPs for this step is  $Nm_1^{[2]}m_2^{[3]}(2m_1^{[3]} - 1)$ . We reshape the gradient into  $\frac{\partial \mathcal{J}}{\partial B_i^{[2]}\text{reshape}(X^{[2]})} \in \mathbb{R}^{m_2^{[3]} \times Nm_1^{[2]}}$ . So, we can get the gradient on  $B_i^{[2]}$  as

$$\frac{\partial \mathcal{J}}{\partial B_i^{[2]}} = \frac{\partial \mathcal{J}}{\partial B_i^{[2]}\text{reshape}(X^{[2]})} \text{reshape}(X^{[2]})^T, \quad (50)$$

with the FLOPs of  $m_2^{[2]}m_2^{[3]}(2Nm_1^{[2]} - 1)$ . The gradient on  $\text{reshape}(X^{[2]})$  is

$$\frac{\partial \mathcal{J}}{\partial \text{reshape}(X^{[2]})} = \sum_{i=1}^{r^{[2]}} B_i^{[2]T} \frac{\partial \mathcal{J}}{\partial B_i^{[2]}\text{reshape}(X^{[2]})}, \quad (51)$$

with the FLOPs  $r^{[2]}(m_1^{[2]}Nm_2^{[2]}(2m_2^{[3]} - 1) + (r^{[2]} - 1)Nm^{[2]}$ . Then we reshape it to get  $\frac{\partial \mathcal{J}}{\partial X^{[2]}} \in \mathbb{R}^{Nm_2^{[2]} \times m_1^{[2]}}$ . Then we calculate the gradient on  $O^{[1]}$  as

$$\frac{\partial \mathcal{J}}{\partial O^{[1]}} = \left( \frac{\partial \mathcal{J}}{\partial X^{[2]}} \right) \odot (\sigma(O^{[1]})), \quad (52)$$

with FLOPs  $Nm^{[2]}$ . Then we can get

$$\frac{\partial \mathcal{J}}{\partial (S^{[1]} \odot A_i^{[1]})} = \left( \frac{\partial \mathcal{J}}{\partial O^{[1]}} \right)^T \text{reshape}(B_i^{[1]}\text{reshape}(X^{[1]})). \quad (53)$$

The FLOPs for this step is  $m_1^{[1]}m_1^{[2]}(2Nm_2^{[2]} - 1)$ . So, the gradients on  $S^{[1]}$  and  $A_i^{[1]}$  are

$$\frac{\partial \mathcal{J}}{\partial S^{[1]}} = \sum_{i=1}^{r^{[1]}} \frac{\partial \mathcal{J}}{\partial (S^{[1]} \odot A_i^{[1]})} \odot A_i^{[1]}, \quad (54)$$

with FLOPs  $r^{[1]}m_1^{[1]}m_1^{[2]} + (r^{[1]} - 1)m_1^{[1]}m_1^{[2]}$ , and

$$\frac{\partial \mathcal{J}}{\partial S^{[1]}} = \frac{\partial \mathcal{J}}{\partial (S^{[1]} \odot A_i^{[1]})} \odot S^{[1]}, \quad (55)$$

with FLOPs  $m_1^{[1]}m_1^{[2]}$ . The gradient on  $\text{reshape}(B_i^{[1]}\text{reshape}(X^{[1]}))$  is

$$\frac{\partial \mathcal{J}}{\partial \text{reshape}(B_i^{[1]}\text{reshape}(X^{[1]}))} = \frac{\partial \mathcal{J}}{\partial O^{[1]}}(S^{[1]} \odot A_i^{[1]}). \quad (56)$$

The FLOPs for this step is  $Nm_1^{[1]}m_2^{[2]}(2m_1^{[2]} - 1)$ . Then we reshape it to get  $\frac{\partial \mathcal{J}}{\partial B_i^{[1]}\text{reshape}(X^{[1]})} \in \mathbb{R}^{m_2^{[2]} \times Nm_1^{[1]}}$ . So, we can get the gradient on  $B_i^{[1]}$  as

$$\frac{\partial \mathcal{J}}{\partial B_i^{[1]}} = \frac{\partial \mathcal{J}}{\partial B_i^{[1]}\text{reshape}(X^{[1]})} \text{reshape}(X^{[1]})^T, \quad (57)$$

with FLOPs  $m_2^{[1]}m_2^{[2]}(2Nm_1^{[1]} - 1)$ .

Therefore, we can get the total FLOPs for the backward pass as

$$\begin{aligned} & Nm^{[3]} + r^{[2]}(m_1^{[2]}m_1^{[3]})(2Nm_2^{[3]} - 1) + r^{[2]}m_1^{[2]}m_1^{[3]} + (r^{[2]} - 1)Nm^{[2]} + r^{[2]}m_1^{[2]}m_1^{[3]} + r^{[2]}Nm_1^{[2]}m_2^{[3]}(2m_1^{[3]} - 1) + \\ & r^{[2]}m_2^{[2]}m_2^{[3]}(2Nm_1^{[2]} - 1) + r^{[2]}(m_1^{[2]}m_2^{[2]})(2m_2^{[3]} - 1) + (r^{[2]} - 1)Nm^{[2]} + Nm^{[2]} + r^{[1]}(Nm_1^{[1]}m_1^{[2]}(2Nm_2^{[2]} - 1) + \\ & r^{[1]}m_1^{[1]}m_1^{[2]} + (r^{[1]} - 1)m_1^{[1]}m_1^{[2]} + r_1m_1^{[1]}m_1^{[2]} + r^{[1]}Nm_1^{[1]}m_2^{[2]}(2m_1^{[2]} - 1) + r^{[1]}m_2^{[1]}m_2^{[2]}(2Nm_1^{[1]} - 1) \\ & = \mathcal{O}\left(N(m^{[2]} + m^{[3]}) + r^{[2]}Nm_1^{[2]}(4m^{[3]} - m_2^{[3]}) + 2r^{[2]}Nm^{[2]}m_2^{[3]} + r^{[1]}Nm_1^{[3]}(4m^{[2]} - m_2^{[2]}) + 2r^{[1]}Nm^{[1]}m_2^{[2]}\right). \end{aligned} \quad (58)$$

□

## **B Computation resource**

we used a server with 64 CPUs of AMD EPYC 7313 16-Core Processor. The server has 8 RTX A5000 GPUs, with 24GB memory for each one. For the experiment with linear model and LeNet, we used only one single GPU. And for the ViT-tiny experiment, we use 2 GPUs at the same time.

## **C Experiment Setting**

To get the linear and LeNet experiment result, *cd* into the folder 'Linear&LeNet' and *python kpd.lenet.py* and *python kpd.one\_layer.py*.

To get the ViT-tiny experiment result, *cd* into the folder 'ViT' and use *bash script/train\_cifar\_kron\_rank4patch4x4.sh*