Supplemental Materials for: Exploring Hidden Dimensions in Parallelizing Convolutional Neural Networks

1. Node and Edge Eliminations

We define node and edge eliminations in Algorithm 1.

Algorithm 1 Node and edge eliminations.

```
function NODEELIMINATION(𝒢)
    if exist a node l_j with a single in-edge e_1 = (l_i, l_j) and a single out-edge e_2 = (l_j, l_k) then
        e' = (l_i, l_k)
        𝒢' = 𝒢 − l_j − e_1 − e_2 + e'
        return 𝒢'
    else
        return 𝒢
    end if
end function

function EDGEELIMINATION(𝒢)
    if exist two edges e_1 = (l_i, l_j) and e_2 = (l_i, l_j) then
        e' = (l_i, l_j)
        𝒢' = 𝒢 − e_1 − e_2 + e'
        return 𝒢'
    else
        return 𝒢
    end if
end function
```

Theorem 1. *Assume $\mathcal{G}'$ = NodeElimination($\mathcal{G}$) and l_j is the eliminated layer. If ${\mathcal{S}_o}'$ is an optimal strategy for $\mathcal{G}'$, then $\mathcal{S}_o = {\mathcal{S}_o}' + \widehat{c_j}$ is an optimal strategy for $\mathcal{G}$, where*

$$\widehat{c_j} = \arg\min_{c_j}\{t_C(n_j, c_j) + t_S(n_j, c_j) + t_X(e_1, c_i, c_j) + t_X(e_2, c_j, c_k)\} \tag{1}$$

Proof. It is equivalent to prove that $t_O(\mathcal{G}, \mathcal{S}_1) \geq t_O(\mathcal{G}, \mathcal{S}_o)$ for any other strategy $\mathcal{S}_1$. We assume layer l_i has parallelization configuration $c_{i1} \in \mathcal{S}_1$. We prove this inequality by using the following path.

$$t_O(\mathcal{G}, \mathcal{S}_1) \geq t_O(\mathcal{G}', \mathcal{S}_1) \tag{2}$$
$$\geq t_O(\mathcal{G}', {\mathcal{S}_o}') \tag{3}$$
$$= t_O(\mathcal{G}, \mathcal{S}_o) \tag{4}$$

Proof of Equation 2. The difference between $t_O(\mathcal{G}, \mathcal{S}_1)$ and $t_O(\mathcal{G}', \mathcal{S}_1)$ is

$$\begin{aligned} &t_O(\mathcal{G}, \mathcal{S}_1) - t_O(\mathcal{G}', \mathcal{S}_1) \\ =&t_C(l_j, c_{j1}) + t_S(l_j, c_{j1}) + t_X(e_1, c_{i1}, c_{j1}) \\ &+ t_X(e_2, c_{j1}, c_{k1}) - t_X(e', c_{i1}, c_{k1}) \end{aligned} \tag{5}$$

This is because all other layers except l_j use the same configurations in $t_O(\mathcal{G}, \mathcal{S}_1)$ and $t_O(\mathcal{G}', \mathcal{S}_1)$, and therefore all cost functions non-related to l_j are eliminated in the subtraction. The remaining parts are l_j, e_1, and e_2, which no longer exist in $\mathcal{G}'$ after node elimination, and e' that is added to $\mathcal{G}'$. Recall that $t_X(e', \cdot, \cdot)$ is defined as follows.

$$\begin{aligned} t_X(e', c_i, c_k) = \min_{c_j}\{&t_C(l_j, c_j) + t_S(l_j, c_j) \\ &+ t_X(e_1, c_i, c_j) + t_X(e_2, c_j, c_k)\} \end{aligned} \tag{6}$$

Combining Equation 5 and 6, we have $t_O(\mathcal{G}, \mathcal{S}_1) \geq t_O(\mathcal{G}', \mathcal{S}_1)$.

Proof of Equation 3. Since ${\mathcal{S}_o}'$ is an optimal strategy for $\mathcal{G}'$, the inequality holds by definition.

Proof of Equation 4. Similarly, the difference between $t_O(\mathcal{G}', {\mathcal{S}_o}')$ and $t_O(\mathcal{G}, \mathcal{S}_o)$ is

$$\begin{aligned} &t_O(\mathcal{G}, \mathcal{S}_o) - t_O(\mathcal{G}', {\mathcal{S}_o}') \\ =&t_C(l_j, \widehat{c_j}) + t_S(l_j, \widehat{c_j}) + t_X(e_1, c_i, \widehat{c_j}) \\ &+ t_X(e_2, \widehat{c_j}, c_k) - t_X(e', c_i, c_k) \end{aligned} \tag{7}$$

This is because $\mathcal{S}_o = {\mathcal{S}_o}' + \widehat{c_j}$, and therefore all cost functions non-related to l_j are eliminated. We can prove Equation 4 by bringing Equation 1 into Equation 7. □

Theorem 2. *Assume $\mathcal{G}'$ = EdgeElimination($\mathcal{G}$), and ${\mathcal{S}_o}'$ is an optimal strategy for $\mathcal{G}'$, then $\mathcal{S}_o = {\mathcal{S}_o}'$ is an optimal strategy for $\mathcal{G}$.*

Proof. We can use the same path to prove this theorem.

Proof of Equation 2. The difference between $t_O(\mathcal{G}, \mathcal{S}_1)$ and $t_O(\mathcal{G}', \mathcal{S}_1)$ is

$$\begin{aligned} &t_O(\mathcal{G}, \mathcal{S}_1) - t_O(\mathcal{G}', \mathcal{S}_1) \\ =&t_X(e_1, c_{i1}, c_{j1}) + t_X(e_2, c_{i1}, c_{j1}) - t_X(e', c_{i1}, c_{j1}) \end{aligned} \tag{8}$$

Recall that $t_X(e', \cdot, \cdot)$ is defined as follows.

$$t_X(e', c_i, c_j) = t_X(e_1, c_i, c_j) + t_X(e_2, c_i, c_j) \tag{9}$$

Combining Equation 8 and 9, we have $t_O(\mathcal{G}, \mathcal{S}_1) = t_O(\mathcal{G}', \mathcal{S}_1)$.

Proof of Equation 3. The inequality holds since ${\mathcal{S}_o}'$ is an optimal strategy for $\mathcal{G}'$.

Proof of Equation 4. The difference between $t_O(\mathcal{G}', {\mathcal{S}_o}')$ and $t_O(\mathcal{G}, \mathcal{S}_o)$ is

$$\begin{aligned} & t_O(\mathcal{G}, \mathcal{S}_o) - t_O(\mathcal{G}', {\mathcal{S}_o}') \\ = & t_X(e_1, c_i, c_j) + t_X(e_2, c_i, c_j) - t_X(e', c_i, c_j) \\ = & 0 \end{aligned} \tag{10}$$

□