# A THE PROOF OF EQ. (8)

Proof. For convolutional Layer (shown in Fig. 4), $z_i^l$ denotes one pixel in the $d^l$ channels of the layer $l$. It connects from $c^l \cdot k^l \cdot k^l$ pixels in the layer $l-1$. In the forward propagation phase, $z_i^l$ is the weighted sum of $c^l \cdot k^l \cdot k^l$ items of $O^{l-1}$ (plus biases, which would be omitted in the following analysis), where the weights are from the filters $W^l$, i.e.

$$z_i^l = \sum_{c^l} \sum_{\cdot k^l \cdot k^l} o_j^{l-1} \cdot w_j^l \tag{11}$$

where $o_j^{l-1} \in O^{l-1}$, $w_j^l \in W^l$. We assume that the elements in $Z$ are mutually independent and share the same distribution, and have the same assumption to $O$ and $W$. We have

$$\mathrm{Var}(Z^l) = c^l \cdot k^l \cdot k^l \mathrm{Var}(O^{l-1})\mathrm{Var}(W^l) \tag{12}$$

Similarly, in the backward propagation phase, we have

$$\mathrm{Var}(\frac{\partial f}{\partial W^l}) = k^l \cdot k^l \mathrm{Var}(\frac{\partial f}{\partial Z^l})\mathrm{Var}(O^{l-1}) \tag{13}$$

$$\mathrm{Var}(\frac{\partial f}{\partial Z^{l-1}}) = k^l \cdot k^l \cdot d^l \mathrm{Var}(\frac{\partial f}{\partial Z^l})\mathrm{Var}(W^l)\mathrm{Var}(\frac{\partial O^{l-1}}{\partial Z^{l-1}}) \tag{14}$$

From the property of ReLU(.) and the same assumption as [38], we have

$$\mathrm{Var}(\frac{\partial O^l}{\partial Z^l}) = \frac{1}{2}, \qquad \mathrm{Var}(O^l) = \frac{1}{2}\mathrm{Var}(Z^l) \tag{15}$$

From Eq. (12)(13)(14)(15), we get

$$\mathrm{Var}(\frac{\partial f}{\partial W^l}) = \frac{k^l \cdot k^l \cdot c^l}{k^{l-1} \cdot k^{l-1} \cdot d^l}\mathrm{Var}(\frac{\partial f}{\partial W^{l-1}})$$

□

# B THE PROOF OF EQ. (9)

For fully-connected layer, the output of layer $l$ is $O^l = \phi(Z^l)$, $Z^l = W^l * O^{l-1} + b^l$. Here $W^l \in \mathbb{R}^{c^l \times d^l}$ is the weight where $c_l$ and $d_l$ are the input and output dimensions respectively. $Z_i^l$ is the weighted sum of $c^l$ items of $Z^{l-1}$ (plus biases, which would be omitted in the following analysis), where the weights are from the filters $W^l$. We suppose that the elements in $Z$ are mutually independent and share the same distribution, and have the same assumption to $O$ and $W$. For the forward propagation phase, we have

$$\mathrm{Var}(Z^l) = c^l \mathrm{Var}(Z^{l-1})\mathrm{Var}(W^l) \tag{16}$$

For the backward propagation phase, we have

$$\mathrm{Var}(\frac{\partial f}{\partial W^l}) = \mathrm{Var}(\frac{\partial f}{\partial Z^l})\mathrm{Var}(O^{l-1}) \tag{17}$$

$$\mathrm{Var}(\frac{\partial f}{\partial Z^{l-1}}) = d^l \mathrm{Var}(\frac{\partial f}{\partial Z^l})\mathrm{Var}(W^l)\mathrm{Var}(\frac{\partial O^{l-1}}{\partial Z^{l-1}}) \tag{18}$$

From Eq. (16)(17)(18)(15), we get

$$\mathrm{Var}(\frac{\partial f}{\partial W^l}) = \frac{c^l}{d^l}\mathrm{Var}(\frac{\partial f}{\partial W^{l-1}})$$

# C THE PROOF OF PROPERTY 1

Proof. For a set with multiple real values, we define the set size to be the difference between the maximum and minimum values of the elements. At each iteration of OptimalAdjustment() in Algorithm 2 , OneStepAdjustment() shrinks the set size containing all the layers' thresholds (from Property 2). Therefore as the iteration goes on, the set size converges to 0. At the end all the layers share the same threshold. □

# D THE PROOF OF PROPERTY 2

Proof. Suppose for two layers named $l_1$ and $l_2$, the gradients follow the same shape distributions with with expectation of 0 denoted as $\mathcal{A}(0, \sigma_1^2)$ and $\mathcal{A}(0, \sigma_2^2)$ (e.g. Gaussian distribution or Laplacian distribution), the numbers of which are $c_1$ and $c_2$ respectively. The sparsity ratios are $\alpha_1$ and $\alpha_2$, and the top-$k$ thresholds are denoted by $th_1$ and $th_2$ with initial values $k_1$ and $k_2$. After execution of Algorithm 2 , the thresholds turn to $th_1'$ and $th_2'$. Here we suppose $k_1 < k_2$.

If the $th_1$ moves from $k_1$ to $k_2$, $th_2$ would move in reverse from $k_2$ to $k_1$ in order to keep the total number of uploaded gradients constant. For $l_1$ and $l_2$, the number of the gradients, the amplitudes of which are between $k_1$ and $k_2$, is $c_1(\alpha_1 - 2(1 - cdf(\frac{k_2}{\sigma_1})))$ and $c_2(2(1-cdf(\frac{k_1}{\sigma_2}))-\alpha_2)$, which are denoted by $m_1$ and $m_2$. If $m_1 < m_2$, when $th_2$ reaches $k_2$, $th_2$ still does not reach $k_1$, i.e. $k_1 < th_2 < k_2$. So if we set $th_1' = \frac{k_1+k_2}{2}$, $th_2'$ satisfies $k_1 < th_2' < k_2$. Now we get

$$\|th_1' - th_2'\| \le \frac{\|th_1 - th_2\|}{2}$$

If $m_1 > m_2$, similarly if we set $th_2' = \frac{k_1+k_2}{2}$, $th_1'$ satisfies $k_1 < th_1' < k_2$. So

$$\|th_1' - th_2'\| \le \frac{\|th_1 - th_2\|}{2} \tag{19}$$

Therefore, the function OneStepAdjustment() in Algorithm 2 reduces the threshold difference by at least half.

□