

A ILLUSTRATION OF THRESHOLD ADJUSTMENT

Fig. 14 illustrates the aim of threshold adjustment of DLS. For any two layers, DLS try to make the top- k thresholds equal by adjusting the sparsity ratios.

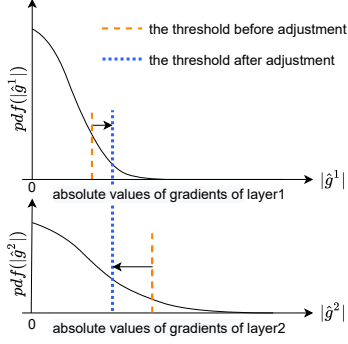


Figure 14: Schematic diagram of threshold adjustment. Given two layers, if the thresholds of them are different, we should adjust the ratios to reduce the larger threshold and increase the smaller one. The ideal state is to adjust the two threshold values to be equal.

B 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.

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

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

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

Similarly, in the backward propagation phase, we have

$$\text{Var}\left(\frac{\partial f}{\partial W^l}\right) = k^l \cdot k^l \text{Var}\left(\frac{\partial f}{\partial Z^l}\right) \text{Var}(O^{l-1}) \quad (13)$$

$$\text{Var}\left(\frac{\partial f}{\partial Z^{l-1}}\right) = k^l \cdot k^l \cdot d^l \text{Var}\left(\frac{\partial f}{\partial Z^l}\right) \text{Var}(W^l) \text{Var}\left(\frac{\partial O^{l-1}}{\partial Z^{l-1}}\right) \quad (14)$$

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

$$\text{Var}\left(\frac{\partial O^l}{\partial Z^l}\right) = \frac{1}{2}, \quad \text{Var}(O^l) = \frac{1}{2} \text{Var}(Z^l) \quad (15)$$

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

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

□

C 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

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

For the backward propagation phase, we have

$$\text{Var}\left(\frac{\partial f}{\partial W^l}\right) = \text{Var}\left(\frac{\partial f}{\partial Z^l}\right) \text{Var}(O^{l-1}) \quad (17)$$

$$\text{Var}\left(\frac{\partial f}{\partial Z^{l-1}}\right) = d^l \text{Var}\left(\frac{\partial f}{\partial Z^l}\right) \text{Var}(W^l) \text{Var}\left(\frac{\partial O^{l-1}}{\partial Z^{l-1}}\right) \quad (18)$$

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

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

D 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. □

E THE PROOF OF PROPERTY 2

PROOF. Suppose for two layers named l_1 and l_2 , the gradients follow the same shape distributions 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 α_1 and α_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

$$\|th'_1 - th'_2\| \leq \frac{\|th_1 - th_2\|}{2}$$

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

$$\|th'_1 - th'_2\| \leq \frac{\|th_1 - th_2\|}{2} \quad (19)$$

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