# Network structure :

$$a = input \cdot W_1 + b_1$$

$$b = ReLu(a)$$

$$c = b \cdot W_2 + b_2$$

$$d = softmax(c)$$

$$Loss = -\frac{1}{N}\sum \log(d_k) + \frac{\alpha}{2}\left(\|W_1\|^2 + \|W_2\|^2\right)$$

# Shape :

$input.shape : (N_{batch}, N_{in})$

$W_1.shape : (N_{in}, N_{hid})$

$a.shape : (N_{batch}, N_{hid})$

$W_2.shape : (N_{hid}, N_{out})$

$c.shape : (N_{batch}, N_{out})$

$d.shape : (N_{batch}, N_{out})$

$\delta_k.shape : (N_{batch}, N_{out})$ ⟶ 每一行仅第 $k$ 个 = 1，ground truth 对应 $d$ 中的 $d_k$（$d_k$ 为一个数值）

Hint:

$$Y = XW \ , \ \frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y}\frac{\partial Y}{\partial x} = \frac{\partial L}{\partial Y} W^T$$

$$Y = XW \ , \ \frac{\partial L}{\partial W} = X^T \frac{\partial L}{\partial Y}$$

Deduction:

$$\frac{\partial Loss}{\partial w_1} = \frac{1}{N}\left( input^T * \frac{\partial Loss}{\partial a} \right) + \alpha \| w_1 \|$$

$$= \frac{1}{N}\left[ input^T * \left( \frac{\partial Loss}{\partial c} * \frac{\partial c}{\partial b} \odot \frac{\partial b}{\partial a} \right) \right] + \alpha \| w_1 \|$$

$$= \frac{1}{N}\left[ input^T * \left( (d - \delta_k) * W_2^T \odot I(a>0) \right) \right] + \alpha \| w_1 \|$$

已经应用了 softmax 梯度，注意到第一层对 $w_1$ 求导，ReLu 的导数是大于 0 而不是大于等于 0

$$\frac{\partial Loss}{\partial b_1} = \frac{\partial Loss}{\partial a} = \frac{1}{N}\left( (d - \delta_k) * W_2^T \odot I(a>0) \right)$$

$$\frac{\partial Loss}{\partial w_2} = \frac{1}{N}\left( b^T * (d - \delta_k) \right) + \alpha \| W_2 \|$$

$$\frac{\partial Loss}{\partial b_2} = \frac{\partial Loss}{\partial c} = \frac{1}{N}(d - \delta_k)$$