X_train shape (60000 x 28x28)

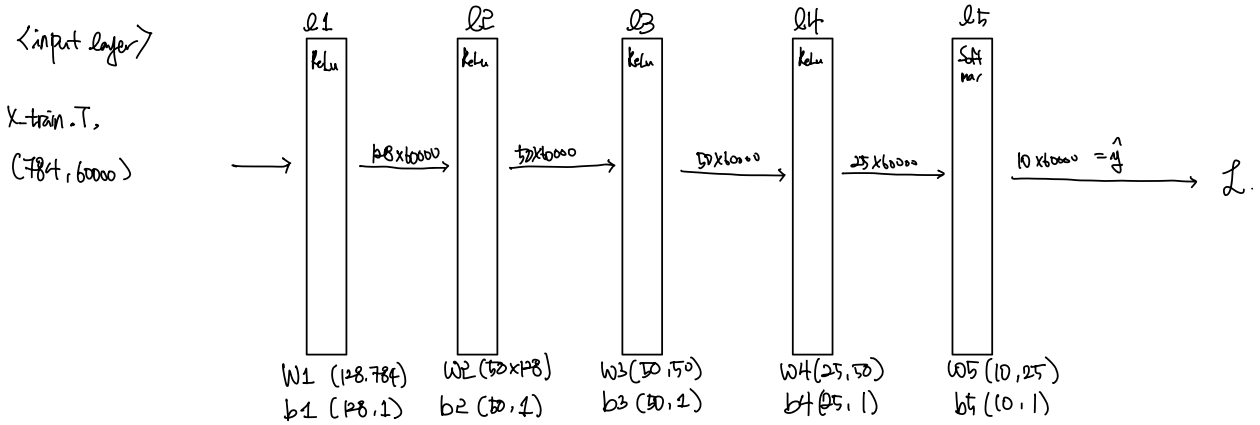
↓ flatten

X_train shape (60000, 784)

X_test shape (10000, 784)

y_train shape (10 x 60000)

y_test shape (10 x 10000)




W1 · X_train.T + b1

(128,784) · (784, 60000) + (128,1)

⇒ (128, 60000) ; Z1.

relu (128, 60000) ; A1.

(10,60000) (10,60000) (10x25) (10x1)

A5, Z5, L. W5, b5

⇐ (1x60000)

$Z_5 = W_5 \cdot A_4 + b_5$

$W_5 = W_5 - \frac{dL}{dW_5} \cdot \alpha$

$\frac{dL}{dW_5} = \frac{dZ_5}{dW_5} \times \frac{dA_5}{dZ_5} \times \frac{dL}{dA_5}$

$L = -\ y\ \ln(A_5)$

$\frac{d}{dA} L = \frac{y}{A_5}$

ex

$Z = \begin{pmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \\ z \end{pmatrix} \rightarrow \sigma(z) = \begin{pmatrix} \frac{e^{z_0}}{e^{z_0} + \square} \\ \vdots \\ e^{z_9} \\ \hline e^{b_4} \square \end{pmatrix}$

$$\frac{dA_5}{dZ_5} = \begin{pmatrix} a_0(1-a_0) & -a_0a_1 & \cdots & -a_0a_9 \\ -a_0a_1 & & & \vdots \\ \vdots & & & \\ a_9a_0 & \cdots & & a_9(1-a_9) \end{pmatrix}$$

$$\frac{dL}{dA_5} = \begin{pmatrix} -\frac{y_{0,1}}{a_{0,1}} & 10000 & \cdots & -\frac{y_{1,10000}}{a_{0,10000}} \\ -\frac{y_2}{a_1} & \cdots & & \vdots \\ \vdots & & & \\ -\frac{y_9}{a_9} & & & -\frac{y_{9,10000}}{a_{9,10000}} \end{pmatrix}$$

$$\frac{dA_5}{dZ_5} \times \frac{dL}{dA_5} = \begin{pmatrix} -y_0(1-a_0) + a_0y_1 + a_0y_2 + \cdots + a_0y_9 \\ -y_1(1-a_1) + a_1 \\ \vdots \end{pmatrix}$$

$$-y_0 + a_0y_0 + a_0(y_1 + \cdots + y_9)$$

$$-y_0 + a_0$$

$$\frac{d}{dZ_5}L = A_5 - y \quad (10 \times 10000)$$

$$A_4 \times (A_5 - y) \quad (25 \times 10000)$$

here is forest. think I should learn matrix more. still don't know

$$\frac{d}{dW_5}\mathcal{L} = (A_5 - y) \times A_4^T \quad (10 \times 25) \times \frac{1}{m} \quad (m = 60000)$$

why

$A_4 \times (A_5 - y)$

⇓

$(A_5 - y) \times A_4^T$.

(I saw a blog about why $A_4 \to A_4^T$, but I don't know if they can apply commutative property in this matrix chain rule.

+) why do I have to divide in 6000?

$$\frac{d}{db_5}\mathcal{L} = \frac{dZ_5}{db_5} \times \frac{dA_5}{dZ_5} \times \frac{d}{dA_5}\mathcal{L}$$

$$= \frac{dZ_5}{db_5} \times (A_5 - y)$$

$$= (A_5 - y)$$

$$\frac{d}{dW_4}\mathcal{L} = \frac{dZ_4}{dW_4} \times \frac{dA_4}{dZ_4} \times \frac{d}{dA_4}\mathcal{L}$$

$$\frac{d}{dA_4}\mathcal{L} = \frac{dZ_5}{dA_4} \times \frac{d}{dZ_5}\mathcal{L}.$$

$$= \frac{dZ_5}{dA_4} \times (A_5 - y) \qquad Z_5 = W_5 \times A_5 + b_1.$$

$$= W_5^T \times (A_5 - y) \to (25, 60000)$$

$$\frac{d}{dW_4}\mathcal{L} = \frac{dZ_4}{dW_4} \times \frac{dA_4}{dZ_4} \times W_5^T \times (A_5 - y)$$

$A_4 = \text{Relu}(Z_4)$ → $(25, 60000)$.

$Z_4 = W_4 \times A_3 + b_4$

$A_3^T \to (60000 \times 50)$. → $A_4 = (25, 60000)$.

$$\frac{d}{db_4}\mathcal{L} = \frac{dZ_4}{db_4} \times \frac{dA_4}{dZ_4} \times \frac{dZ_5}{dA_4} \times dZ_5.$$

1

$$\frac{d}{dw_3} L = \frac{dZ_3}{dw_3} \times \frac{dA_3}{dZ_3} \times \frac{d}{dA_3} L$$

$$\frac{d}{dA_3} L = \frac{dZ_4}{dA_3} \times \frac{d}{dZ_4} L \qquad Z_4 = W_4 \cdot A_3 + b_4$$

$\frac{dZ_4}{dA_3}$ : $W_4^T$ (10 X 4)

$\frac{d}{dZ_4} L$ : (25, 60000)

$$\frac{d}{dw_3} L = \frac{dZ_3}{dw_3} \times \frac{dA_3}{dZ_3} \times \frac{d}{dA_3} L$$

$\frac{dZ_3}{dw_3}$ : $A_2^T$ (60000 X 50)

$\frac{dA_3}{dZ_3}$ : (50 X 60000)

$\frac{d}{dA_3} L$ : (50 X 60000)

$$Z_3 = W_3 X A_2 + b_3$$

Back propagation.

① $\frac{\partial \mathcal{L}}{\partial W_5}$

$\frac{\partial Z_5}{\partial W_5} \times \frac{\partial}{\partial Z_5}\mathcal{L} = np.dot(A_5 - Y, A_4^T) = dW_5, (10 \times 25)$.

$\frac{\partial Z_5}{\partial W_5}$ → $A_4^T$ $(60000 \times 25)$

$\frac{\partial}{\partial Z_5}\mathcal{L}$ → $A_5 - Y$, $(10 \times 60000)$

cache $dZ_5$ $(10 \times 60000)$

② $\frac{\partial \mathcal{L}}{\partial W_4}$

$\frac{\partial Z_4}{\partial W_4} \times \frac{\partial}{\partial Z_4}\mathcal{L}$

$\frac{\partial}{\partial Z_4}\mathcal{L} = \frac{\partial A_4}{\partial Z_4} \times \frac{\partial Z_5}{\partial A_4} \times \frac{\partial}{\partial Z_5}\mathcal{L} = Relu\,back(np.dot(W_5^T, dZ_5))$

$\frac{\partial A_4}{\partial Z_4}$ → Relu back $(25, 60000)$

$\frac{\partial Z_5}{\partial A_4}$ → $W_5^T$ $25 \times 10$

$\frac{\partial}{\partial Z_5}\mathcal{L}$ $(10 \times 60000)$

→ $dZ_4$, $(25, 60000)$. cache.

$\frac{\partial Z_4}{\partial W_4} = A_3^T$, $(60000, 50)$

$dW_4 = np.dot(dZ_4, X_3^T)$

③ $\frac{\partial}{\partial W_3}\mathcal{L}$

$\frac{\partial Z_3}{\partial W_3} \times \frac{\partial}{\partial Z_3}\mathcal{L}$.

$\frac{\partial}{\partial Z_3}\mathcal{L} = \frac{\partial A_3}{\partial Z_3} \times \frac{\partial Z_4}{\partial A_3} \times dZ_4 = Relu\,back(np.dot(W_4^T, dZ_4))$

$\frac{\partial Z_4}{\partial A_3}$ → $W_4^T$

→ $dZ_3$ $(50 \times 60000)$, cache.

$dW_3 = np.dot(dZ_3, A_2^T)$

④ $\frac{\partial}{\partial W_2} L$

$\frac{\partial Z_2}{\partial W_2} \times \frac{\partial}{\partial Z_2} L$

$\frac{\partial}{\partial Z_2} L = \frac{\partial A_2}{\partial Z_2} \times \frac{\partial Z_3}{\partial A_2} \times dZ_3 = \text{Relu\_back}(np.dot(W_3^T, dZ_3))$

$W_3^T$

$\to dZ_2, (50 \times 60000)$, cache.

$dW_2 = np.dot(dZ_2, A_1^T)$

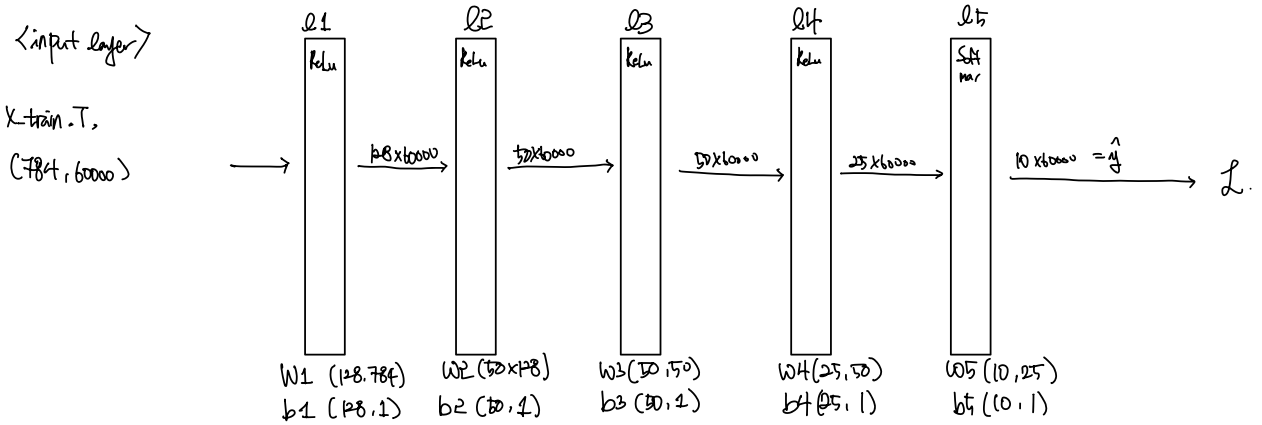
⑤ $\frac{\partial}{\partial W_1} L$

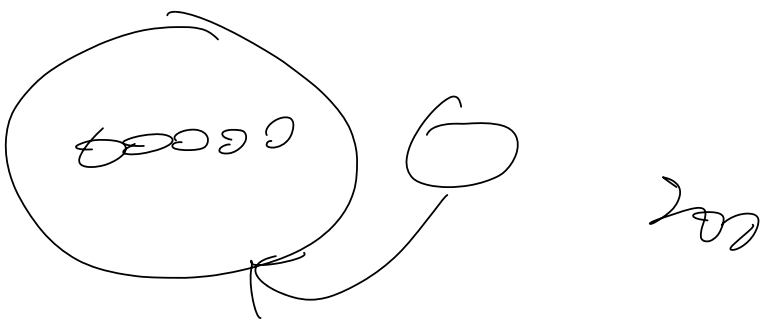
$\frac{\partial Z_1}{\partial W_1} \times \frac{\partial}{\partial Z_1} L$

$\frac{\partial}{\partial Z_1} L = \frac{\partial A_1}{\partial Z_1} \times \frac{\partial Z_2}{\partial A_1} \times dZ_2 = \text{Relu\_back}(np.dot(dZ_2, W_2^T))$

$\to dZ_1, (128 \times 60000)$ cache.

$dW_1 = np.dot(dZ_1, X\_train)$



$(128 \times 784) \times (784 \times 1) = 128 \times 1$

$(50 \times 128) \times (128 \times 1) = 50 \times 1$

$50 \times 50 \times 50 \times 1 = 50 \times 1$

$25 \times 50 \times 50 \times 1 = 25 \times 1$

$10 \times 25 \times 25 \times 1 = 10 \times 1$

$(10 \times 1) - (10 \times 1) = (10 \times 1)$

$(10 \times 1) \times (1 \times 25) = 10 \times 25$