#1.

$$L = \frac{1}{2} \| W^T W x - x \|^2$$

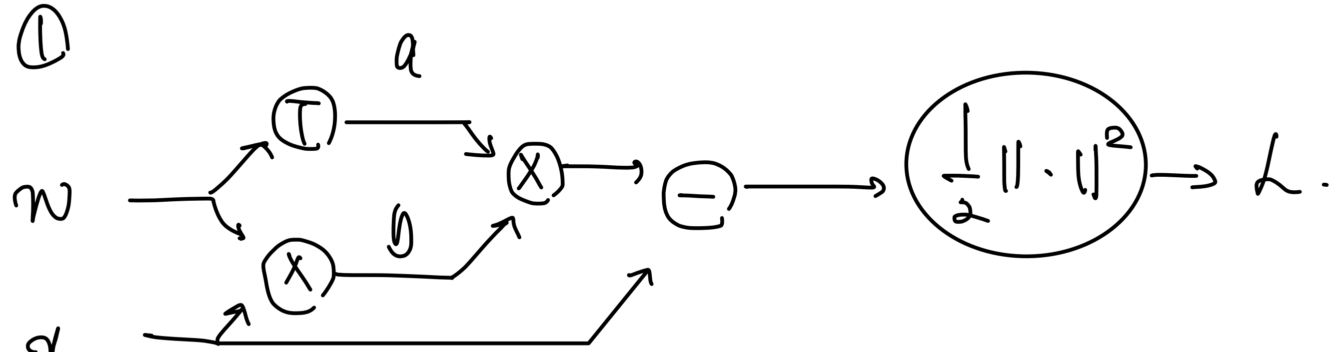
(a)

The transformation $W^T W x$ first encodes information through Wx and then decodes it through W^T.
The loss function L measures how well the recovered information $W^T W x$ preserves the original input x.
By minimizing the loss, the model learns to retain important features of x during encoding and decoding.

(b)

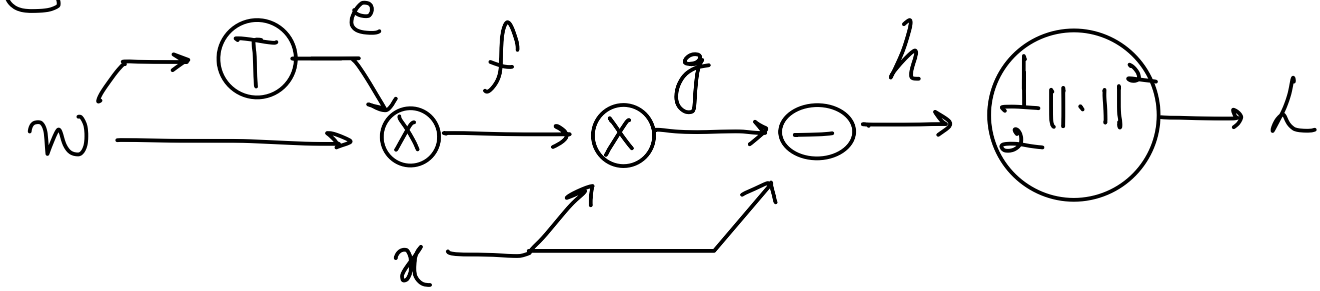
$$\mathcal{L} = \frac{1}{2}\|W^T W x - x\|^2$$

①



OR

②

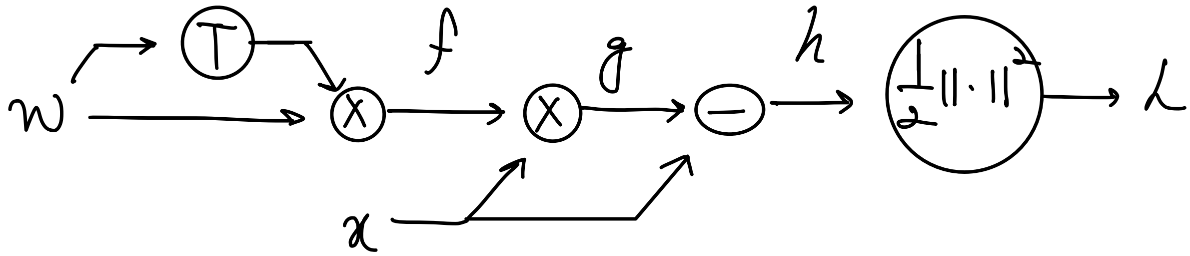


(c)

On the graph (b)-1, W contributes to two paths, a and b
So, its gradient contributions must be accumulated accordingly
by the total derivative rule such that

$$\nabla_W L = \frac{dL}{da}\cdot\frac{da}{dW} + \frac{dL}{db}\cdot\frac{db}{dW}.$$

(d) $L = \frac{1}{2}\|W^TWx - x\|^2$



$x \in \mathbb{R}^n$ $\quad e \in \mathbb{R}^{n \times m}$ $\quad g \in \mathbb{R}^n$ $\quad L \in \mathbb{R}$

$w \in \mathbb{R}^{m \times n}$ $\quad f \in \mathbb{R}^{n \times n}$ $\quad h \in \mathbb{R}^n$

$f = W^TW$ $\qquad L = \frac{1}{2}\|h\|^2$

$g = f \cdot x$ $\qquad h = g - x$

$$\frac{dL}{dh} = h$$

$$\frac{dL}{dg} = \frac{dL}{dh} \cdot \frac{dh}{dg} = h \cdot I = h$$

$$\frac{dL}{df} = \frac{dL}{dg} \cdot \frac{dg}{df} = h \cdot x^T \quad [\in \mathbb{R}^{n \times n}]$$

$$\frac{dL}{dW} = \frac{df}{dW} \cdot \frac{dL}{df} = 2W \cdot (h \cdot x^T) \quad [\in \mathbb{R}^{m \times n}]$$

$$\Rightarrow \nabla_W L = 2W \cdot (W^TWx - x)x^T$$

#2

I am a C147 Student.

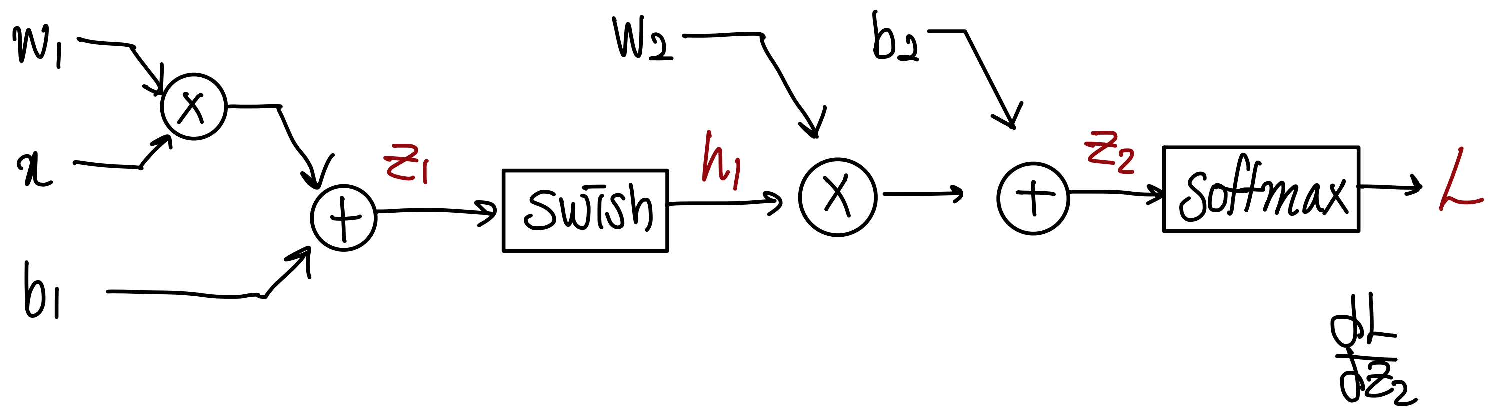
#3.

$$\boxed{x} \to \boxed{z_1 = W_1 x + b_1} \to \boxed{h_1 = \text{swish}(z_1)} \to \boxed{z_2 = W_2 h_1 + b_2} \to \boxed{\text{softmax}(z_2)} \to \boxed{L}$$

$H \times 1$

(a)

$x \in \mathbb{R}^D$, $z_1 \in \mathbb{R}^H$, $W_1 \in \mathbb{R}^{H\times D}$, $b_1 \in \mathbb{R}^H$, $h_1 \in \mathbb{R}^H$, $z_2 \in \mathbb{R}^C$, $W_2 \in \mathbb{R}^{C\times H}$, $b_2 \in \mathbb{R}^C$



(b) $\nabla_{W_2} L$, $\nabla_{b_2} L$,,

$$z_2 = W_2 h_1 + b_2$$

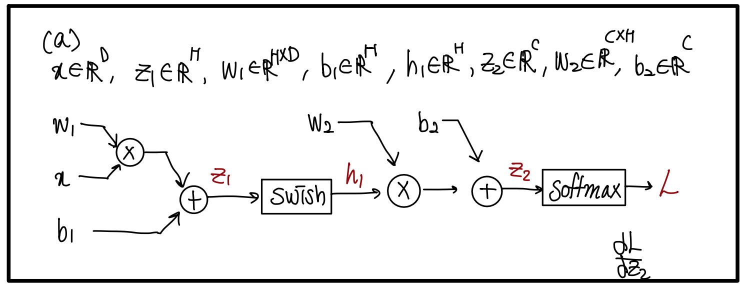
$$\frac{dL}{db_2} = \frac{dL}{dz_2} \cdot \frac{dz_2}{db_2} = \frac{dL}{dz_2}$$

$$\frac{dL}{dW_2} = \frac{dL}{dz_2} \cdot \frac{dz_2}{dW_2} = \frac{dL}{dz_2} \cdot h_1^T \quad [\mathbb{R}^{C\times H}]$$

$$\Rightarrow \quad \nabla_{W_2} L = \frac{dL}{dz_2} \cdot h_1^T, \quad \nabla_{b_2} L = \frac{dL}{dz_2}$$

(c) $\nabla_{W_1} L$, $\nabla_{b_1} L$ //

$$\boxed{x} \rightarrow \boxed{z_1 = W_1 x + b_1} \rightarrow \boxed{h_1 = \mathrm{Swish}(z_1)} \rightarrow \boxed{z_2 = W_2 h_1 + b_2} \rightarrow \boxed{\mathrm{softmax}(z_2)} \rightarrow \boxed{L}$$



$$h_1 = z_1 \cdot \sigma(z_1)$$

$$\frac{dh_1}{dz_1} = \sigma(z_1) + z_1 \sigma(z_1)(1 - \sigma(z_1)) \quad [\in \mathbb{R}^{C \times H}]$$

$$\frac{dL}{dh_1} = \frac{dz_2}{dh_1} \cdot \frac{dL}{dz_2} = W_2^T \cdot \frac{dL}{dz_2} \quad [\in \mathbb{R}^H]$$

$$\frac{dL}{dz_1} = \frac{dh_1}{dz_1} \cdot \frac{dL}{dh_1} = [\sigma(z_1) + z_1 \sigma(z_1)(1 - \sigma(z_1))] \cdot \left(W_2^T \cdot \frac{dL}{dz_2}\right) \quad [\in \mathbb{R}^C]$$

$$\frac{dL}{dW_1} = \frac{dL}{dz_1} \cdot \frac{dz_1}{dW_1} = \frac{dL}{dz_1} \cdot x^T \quad [\in \mathbb{R}^{H \times D}]$$

$$\frac{dL}{db_1} = \frac{dL}{dz_1} \cdot \frac{dz_1}{db_1} = \frac{dL}{dz_1} \quad [\in \mathbb{R}^C]$$

$$\Rightarrow \nabla_{W_1} L = [\sigma(z_1) + z_1 \sigma(z_1)(1 - \sigma(z_1))] \cdot x^T$$

$$\nabla_{b_1} L = [\sigma(z_1) + z_1 \sigma(z_1)(1 - \sigma(z_1))]$$

, where $\sigma(K)$ is the sigmoid activation function.