① W_2, b_2：隐藏层 $\xrightarrow{softmax}$ 输出层($\hat{y}$) $\longrightarrow$ 监督数据 y

$\boxed{W_2}$ 令 n 为特征个数，K 为目标分类个数（在本题中，$n=256$，$K=10$），则对于一个 $n\times1$ 的输入样本向量 $x^{(1)}=(x_1^{(1)}\cdots x_n^{(1)})^T$；一个 $n\times K$ 的权重矩阵 W_2，及一个 $K\times1$ 的偏置向量 b_2，$z=W^Tx^{(1)}+b$，$z_j=\sum_{i=1}^{n}W_{j,i}\,x_i^{(1)}+b_j$

$\hat{y}$ 与 y 均为 $K\times1$ 的向量，$\hat{y}_j=softmax(z)_j=\dfrac{exp(z_j)}{\sum_{q=1}^{K}exp(z_q)}$

不考虑正则化时，$Loss(\hat{y},y)=-\sum y_j\log(\hat{y}_j)=-\log(\hat{y}_t)$（$t$ 为正确分类）

$$\frac{\partial L}{\partial z_t}=\frac{\partial L}{\partial \hat{y}_t}\cdot\frac{\partial \hat{y}_t}{\partial z_t}$$，其中，$\dfrac{\partial L}{\partial \hat{y}_t}=-\dfrac{1}{\hat{y}_t}$

对 $$\frac{\partial \hat{y}_t}{\partial z_t}=\frac{\partial softmax(z)_t}{\partial z_t}=\frac{exp(z_t)\sum exp(z_q)-exp(z_t)^2}{(\sum exp(z_q))^2}=\hat{y}_t(1-\hat{y}_t)$$

1° $j\neq t$

$$\frac{\partial L}{\partial z_t}=-\frac{1}{\hat{y}_t}\cdot\hat{y}_t(1-\hat{y}_t)=\hat{y}_t-1$$

2° $j=t$

$$\frac{\partial \hat{y}_t}{\partial z_j}=exp(z_t)\frac{\partial\left(\frac{1}{\sum exp(z_q)}\right)}{\partial z_j}=-\hat{y}_t\cdot\hat{y}_j$$

$$\frac{\partial L}{\partial z_t}=-\frac{1}{\hat{y}_t}\cdot(-\hat{y}_t\cdot\hat{y}_j)=\hat{y}_j$$

$$\therefore\ \frac{\partial L}{\partial z}=[\hat{y}_1,\hat{y}_2\ \cdots\ \hat{y}_{t-1},\ \cdots\ \hat{y}_K]=\hat{y}-y$$

$$\therefore\ \frac{\partial L}{\partial W_i}=\frac{\partial L}{\partial z}\cdot\frac{\partial z}{\partial W_i}=(\hat{y}-y)\,x_i^{(1)}$$

即 $\dfrac{\partial L}{\partial W_2}=(\hat{y}-y)x^{(1)}$，其中 $\hat{y}$ 是最终输出，y 是监督数据，$x^{(1)}$ 是隐藏层到监督层的输入

此时将样本量与 L_2 正则项纳入考虑，设样本量为 m

$$\frac{\partial L}{\partial W_2} = (\hat{Y}-Y)X^{(1)}/m + \frac{\partial L}{\partial C_{L2}} = (\hat{Y}-Y)X^{(1)} + \frac{\partial L}{\partial\left(\frac{\lambda}{2m}\left(\sum_{i=1}^{L} W_i^T W_i\right)\right)}$$

$$= (\hat{Y}-Y)X^{(1)}/m + \frac{\lambda}{m} W_2$$

$\boxed{b_2}$ $$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial b_2} = \frac{\partial L}{\partial z} = (\hat{Y}-Y)/m$$

② W_1, b_1 : 输入层 $\xrightarrow{ReLu}$ 隐藏层

$\boxed{W_1}$ 设输入向量为 $X^{(0)} = (X_1^{(0)}, \cdots X_n^{(0)})^T$，权重矩阵 W_1 为 $n\times n$，偏置向量 b_1 为 $n\times 1$， $h = W^T x^{(0)} + b$ ，$h_j = \sum_{i=1}^{n} W_{j,i} X_i^{(0)} + b_j$

$x^{(1)}$ 是 h 经 ReLu 函数后的输出， $X_j^{(1)} = \max(0, h_j)$

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial h} \cdot \frac{\partial h}{\partial W_1} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial X^{(1)}} \cdot \frac{\partial X^{(1)}}{\partial h} \cdot \frac{\partial h}{\partial W_1}$$

$$\frac{\partial L}{\partial z} = \hat{Y} - Y, \quad \frac{\partial z}{\partial X^{(1)}} = W_2^T$$

$$\frac{\partial X^{(1)}}{\partial h_j} = \begin{cases} 0, & 0 \geq h_j \\ 1, & 0 < h_j \end{cases} \qquad \frac{\partial h}{\partial W_1} = X^{(0)}$$

$\therefore$ $\frac{\partial L}{\partial W_1} = ((\hat{Y}-Y) W_2^T X^{(0)\star})/m$，其中对 $X^{(0)\star}$ 中的每一个样本 $X_j^{(0)\star}$，

$$X_{j,i}^{(0)\star} = \begin{cases} 0, & 0 \geq h_i = \sum_1^n W_{i,k} X_{j,i}^{(0)} + b_i \\ X_{j,i}^{(0)}, & 0 < h_i = \sum_1^n W_{i,k} X_{j,i}^{(0)} + b_i \end{cases}$$

加上正则项的梯度后， $\frac{\partial L}{\partial W_{1,i}} = ((\hat{Y}-Y) W_2^T X^{(0)\star})/m + \frac{\lambda}{m} W_1$

（$X^{(0)\star}$ 如上定义）

$\boxed{b_1}$ $\dfrac{\partial L}{\partial b_1} = \dfrac{\partial L}{\partial z} \cdot \dfrac{\partial z}{\partial X^{(1)}} \cdot \dfrac{\partial X^{(1)}}{\partial h} \cdot \dfrac{\partial h}{\partial b_1}$

$= ((\hat{Y} - Y) \cdot W_2^T \cdot 1^*) / m$, 其中 $1^*_i = \begin{cases} 0, & 0 \geq h_i \\ 1, & 0 < h_i \end{cases}$