$$\hat{y}_i = \text{softmax}(z) = \frac{e^{z_i}}{\sum e^{z_i}} \quad [\text{logits of } c \text{ classes}]$$

$$z = W_L^T h + b_L$$

$$h = g(W_1^T x + b_1)$$

Loss function

$$L(y, \hat{y}) = -\sum_{i=1}^{c} y_i \cdot \log(\hat{y}_i)$$

$y \rightarrow$ one hot encoding

$y_i \rightarrow$ predicted prob

compute derivate

$$\frac{dL}{d\hat{y}_i} = -\frac{y_i}{\hat{y}_i}$$

$$\frac{\partial L}{\partial z_i} = \sum_{k=1}^{c} \frac{\partial L}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial z_i}$$

we will derivate $\hat{y}_k = \frac{e^{z_k}}{\sum e^{z_k}}$

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1-\hat{y}_i) & i = j \\ -\hat{y}_i \hat{y}_j & i \neq j \end{cases}$$

combine $\frac{\partial L}{\partial \hat{y}_i}$, $\frac{\partial \hat{y}_i}{\partial z_j}$ $\left[\frac{\partial L}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}\right]$

$$i = j \rightarrow \frac{\partial L}{\partial z_j} = \hat{y}_i(1-\hat{y}_i) \cdot \left(-\frac{y_i}{\hat{y}_i}\right) = \hat{y}_i - y_i$$

$$i \neq j \rightarrow \frac{\partial L}{\partial z_j} = -\hat{y}_i y_j \cdot \left(-\frac{y_i}{\hat{y}_i}\right) = 0$$