Derive Gradient ~~Descent~~ For $\boxed{\rightarrow \text{softmax} \rightarrow \text{cross entropy} \rightarrow}$

1. $Z \rightarrow \boxed{\text{softmax function}} \rightarrow \text{softmax\_out}$

$$\begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix} \rightarrow \text{softmax\_out}_i = \frac{e^{z_i}}{\sum_{j=0}^{k} e^{z_j}} \rightarrow \begin{bmatrix} \text{softmax\_out}_0 \\ \text{softmax\_out}_1 \\ \text{softmax\_out}_2 \\ \vdots \\ \text{softmax\_out}_k \end{bmatrix}$$

$$\frac{\partial\, \text{softmax\_out}_n}{\partial z_m} = \frac{\partial}{\partial z_m} e^{z_n}\left(e^{z_0} + e^{z_1} + \cdots + e^{z_k}\right)^{-1}$$

let $e^{z_0} + e^{z_1} + \cdots + e^{z_k} = S$

$$= \begin{cases} e^{z_m}(S)^{-1} + e^{z_m}\cdot(-1)(S)^{-2} e^{z_m}, & \text{when } m = n \\ (-1)e^{z_n}(S)^{-2} e^{z_m}, & \text{when } m \neq n \end{cases}$$

$$= \begin{cases} e^{z_m}(S)^{-1}\left(1 - e^{z_m}(S)^{-1}\right), & \text{when } m = n \\ -\left(e^{z_n}S^{-1}\right)\left(e^{z_m}S^{-1}\right) & \text{when } m \neq n \end{cases}$$

$$= \begin{cases} \text{softmax\_out}_m\left(1 - \text{softmax\_out}_m\right) & \text{when } m = n \\ -\left(\text{softmax\_out}_n\right)\left(\text{softmax\_out}_m\right) & \text{when } m \neq n \end{cases}$$

2. $\underset{OUT}{SOFTMAX}$ $\quad Y$

$$\longrightarrow \boxed{\text{cross entropy}} \longrightarrow L$$

$$\begin{bmatrix} softmax\_out_0 \\ softmax\_out_1 \\ \vdots \\ softmax\_out_k \end{bmatrix} \quad \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix}$$

$$\longrightarrow \boxed{-\sum_{i=0}^{k} y_i \log(softmax\_out_i)} \longrightarrow L$$

$$\frac{\partial L}{\partial\, softmax\_out_m} = -(y_m)\left(\frac{1}{softmax\_out_m}\right)$$

3. Combine 1. and 2.

$$\frac{\partial L}{\partial z_m} = \frac{\partial L}{\partial\, SOFTMAX\_OUT} * \frac{\partial\, SOFTMAX\_OUT}{\partial z_m}$$

$$= \sum_{n=0}^{k} \frac{\partial L}{\partial\, softmax\_out_n} * \frac{\partial\, softmax\_out_n}{\partial z_m}$$

assume $y_j = 1$ for $j = l$ else $y_j = 0$ recall that Y is a one hot vector

$$= -y_l\left(\frac{1}{softmax\_out_l}\right) \times \frac{\partial\, softmax\_out_l}{\partial z_m}$$

$$= \begin{cases} -y_m\left(\frac{1}{\cancel{softmax\_out_m}}\right)\left(\cancel{softmax\_out_m}\right)(1 - softmax\_out_m) & \text{when } l = m \\ +y_l\left(\frac{1}{\cancel{softmax\_out_l}}\right)\left(\cancel{softmax\_out_l}\right)(softmax\_out_m) & \text{when } l \neq m \end{cases}$$

$$= softmaxout_m - y_m$$