Cross entropy gradient $\longrightarrow$

$$\mathcal{L} = \frac{1}{m} \sum_{b=1}^{m} \left(- \sum_{p=1}^{n_y} y_{pb} \cdot \overbrace{\log \hat{y}_{pb}}^{g_{pb}} \right)$$

$$g = \log(\hat{y}) \longrightarrow \frac{\partial g_{ke}}{\partial \hat{y}_{ij}} = \frac{1}{\hat{y}_{ke}} \delta_{k,i} \delta_{e,j}$$

$$\frac{\partial \mathcal{L}}{\partial g_{ij}} = \frac{1}{m} \left(- y_{ij} \cdot 1 \right) = - \frac{1}{m} y_{ij}$$

$$\frac{\partial \mathcal{L}}{\partial \hat{y}_{ij}} = \sum_{k=1}^{n_y} \sum_{l=1}^{m} \frac{\partial \mathcal{L}}{\partial g_{ke}} \cdot \frac{\partial g_{ke}}{\partial \hat{y}_{ij}} = \sum_{k=1}^{n_y} \sum_{e=1}^{m} \left(- \frac{1}{m} y_{ke} \right) \cdot \frac{1}{\hat{y}_{ke}} \delta_{k,i} \delta_{e,j} =$$

$$= - \frac{1}{m} \frac{y_{ij}}{\hat{y}_{ij}}$$

- Softmax $\longrightarrow$

$$\frac{\partial \mathcal{L}}{\partial z_{ij}} = \sum_{k=1}^{n_y} \sum_{l=1}^{m} \frac{\partial \mathcal{L}}{\partial \hat{y}_{ke}} \cdot \frac{\partial \hat{y}_{ke}}{\partial z_{ij}} = \sum_{k=1}^{n_y} \sum_{l=1}^{m} \left(- \frac{1}{m} \frac{y_{ke}}{\cancel{\hat{y}_{ke}}} \right) \delta_{e,j} \left[\delta_{k,i} \cancel{\hat{y}_{ke}} (1 - \hat{y}_{ke}) - \hat{y}_{ie} \cancel{\hat{y}_{ke}} \right.$$

$$\left. (1 - \delta_{k,i}) \right] \delta_{e,j} = \sum_{k=1}^{n_y} \left(- \frac{1}{m} y_{kj} \right) \left[\delta_{k,i} (1 - \hat{y}_{kj}) - \hat{y}_{ij} (1 - \delta_{k,i}) \right]$$

$$= \left(- \frac{1}{m} y_{ij} \right) (1 - \cancel{\hat{y}_{ij}}) + \sum_{k=1}^{n_y} \left(+ \frac{1}{m} y_{kj} \right) \hat{y}_{ij} - \cancel{\left(\frac{1}{m} y_{ij} \cdot \hat{y}_{ij} \right)} =$$

$$= - \frac{1}{m} y_{ij} + \frac{1}{m} \hat{y}_{ij} \underbrace{\sum_{k=1}^{n_y} y_{kj}}_{=1} = \frac{1}{m} \left(\hat{y}_{ij} - y_{ij} \right)$$

Softmax gradient $\longrightarrow$

$$A_{a,b} = \frac{e^{z_{ab}}}{\sum_{k=1}^{n_k} e^{z_{kb}}} \quad ; \quad h(x) = g(x)/f(x) \longrightarrow h'(x) = \frac{g'(x) f(x) - g(x) f'(x)}{f^2(x)}$$

$$\frac{\partial A_{a,b}}{\partial z_{i,j}} = \frac{\left(e^{z_{ab}}\right)' \sum_{k=1}^{n_k} e^{z_{kb}} - e^{z_{ab}} \left(\sum_{k=1}^{n_k} e^{z_{kb}}\right)'}{\left(\sum_{k=1}^{n_k} e^{z_{kb}}\right)^2} = \left| 0 \text{ if } j \neq b \right| =$$

$$= \frac{\left(e^{z_{ab}}\right)' \sum_{k=1}^{n_k} e^{z_{kb}} - e^{z_{ab}} \left(\sum_{k=1}^{n_k} e^{z_{kb}}\right)'}{\left(\sum_{k=1}^{n_k} e^{z_{kb}}\right)^2} \delta_{j,b} =$$

$$\underline{i = a} \Rightarrow \frac{e^{z_{ab}}}{\sum_{k=1}^{n_k} e^{z_{kb}}} - \frac{e^{z_{ab}} \cdot e^{z_{ab}}}{\left(\sum_{k=1}^{n_k} e^{z_{kb}}\right)^2} = A_{ab} - A_{ab}^2$$

$$i \neq a \Rightarrow 0 - \frac{e^{z_{ab}} e^{z_{ib}}}{\left(\sum_{k=1}^{n_k} e^{z_{kb}}\right)^2} = - A_{ab} A_{ib}$$

$$\frac{\partial A_{ab}}{\partial z_{ij}} = \left[(A_{ab} - A_{ab}^2)\, \delta_{i,a} - A_{ab} A_{ib} (1 - \delta_{i,a}) \right] \delta_{jb}$$