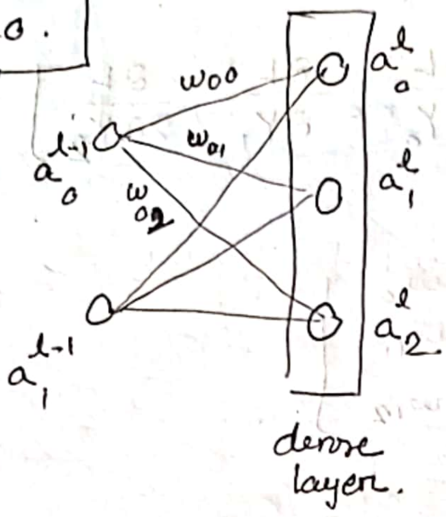
a^{l}_{0} = activation for layer l and node 0.

w_{ij} = weight of edge between ith node of previous layer and jth node of current layer.

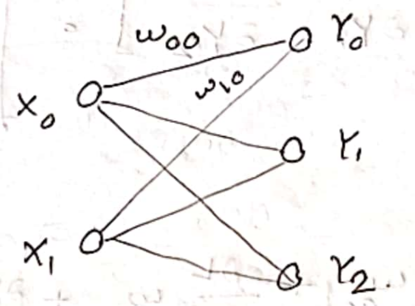


output size = 3
input size = 2.

$$\text{weight matrix} = \begin{bmatrix} w_{00} & w_{01} & w_{02} \\ w_{10} & w_{11} & w_{12} \end{bmatrix}$$

$$\text{bias matrix} = \begin{bmatrix} b_0 & b_1 & b_2 \end{bmatrix}$$

we can interpret like this:



$$Y_0 = w_{00}x_0 + w_{10}x_1 + b_0$$

$$Y_1 = w_{01}x_0 + w_{11}x_1 + b_1$$

$$Y_2 = w_{02}x_0 + w_{12}x_1 + b_2$$

$$\begin{bmatrix} Y_0 & Y_1 & Y_2 \end{bmatrix} = \begin{bmatrix} X_0 & X_1 \end{bmatrix} \begin{bmatrix} w_{00} & w_{01} & w_{02} \\ w_{10} & w_{11} & w_{12} \end{bmatrix} + \begin{bmatrix} b_0 & b_1 & b_2 \end{bmatrix}$$

$$\frac{dY}{dX} = W^T$$

$$\text{grad output} = \frac{\partial L}{\partial Y} = \left[\frac{\partial L}{\partial Y_0} \quad \frac{\partial L}{\partial Y_1} \quad \frac{\partial L}{\partial Y_2}\right]$$

$$w^T = \begin{bmatrix} w_{00} & w_{10} \\ w_{01} & w_{11} \\ w_{02} & w_{12} \end{bmatrix}.$$

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \cdot w^T$$

$$= \left[\frac{\partial L}{\partial Y_0} \quad \frac{\partial L}{\partial Y_1} \quad \frac{\partial L}{\partial Y_2}\right] \cdot \begin{bmatrix} w_{00} & w_{10} \\ w_{01} & w_{11} \\ w_{02} & w_{12} \end{bmatrix}$$

$$= \left[\frac{\partial L}{\partial Y_0} \times w_{00} + \frac{\partial L}{\partial Y_1} w_{01} + \frac{\partial L}{\partial Y_2} w_{02} \quad \frac{\partial L}{\partial Y_0} \times w_{10} \cdots\right]$$

$$= \left[\frac{\partial L}{\partial X_0} \quad \frac{\partial L}{\partial X_1}\right]$$

So, ~~grad input~~ $= \frac{\partial L}{\partial X_0} = \frac{\partial L}{\partial Y_0} \underbrace{\frac{\partial Y_0}{\partial X_0}}_{w_{00}} + \frac{\partial L}{\partial Y_1} \underbrace{\frac{\partial Y_1}{\partial X_0}}_{w_{01}} + \frac{\partial L}{\partial Y_2} \underbrace{\frac{\partial Y_2}{\partial X_0}}_{w_{02}}$

thus this is correct. ~~$= \frac{\partial L}{\partial Y_0} \cdot w_{00}$~~

[effect of everyone will be added]

$\therefore$ grad input $= \begin{bmatrix} \frac{\partial L}{\partial X_0} & \frac{\partial L}{\partial X_1} \end{bmatrix}$

this value will be further back propagated.

in the same way,

grad weight, $\frac{\partial L}{\partial w} = X^T \frac{\partial L}{\partial Y}$

$$= \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \begin{bmatrix} \frac{\partial L}{Y_0} & \frac{\partial L}{Y_1} & \frac{\partial L}{Y_2} \end{bmatrix}$$

$$= \begin{bmatrix} x_0 \frac{\partial L}{Y_0} & x_0 \frac{\partial L}{Y_1} & x_0 \frac{\partial L}{Y_2} \\ x_1 \frac{\partial L}{Y_0} & x_1 \frac{\partial L}{Y_1} & x_1 \frac{\partial L}{Y_2} \end{bmatrix}$$

for the first element of the matrix

$Y_0 = w_{00} x_0 + b.$

$$\therefore \frac{\partial Y_0}{\partial w_{00}} = x_0.$$

$$\therefore \frac{\partial L}{\partial w_{00}} = \cancel{\frac{\partial L}{\partial x}} \frac{\partial Y_0}{\partial w_{00}} \cdot \frac{\partial L}{\partial Y_0}$$

$$= x_0 \frac{\partial L}{\partial Y_0}$$

every calculation prior this page was for ONE SAMPLE for ease

$$\frac{\partial L}{\partial b} = \left[\frac{\partial L}{\partial b_0} \quad \frac{\partial L}{\partial b_1} \quad \frac{\partial L}{\partial b_2} \right]$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial Y} \cdot \frac{\partial Y}{\partial b} = \frac{\partial L}{\partial Y}.$$

Y_{ij} = output of ith node for jth sample

$\therefore$ for ~~one~~ two ~~all~~ samples

$$\frac{\partial L}{\partial Y} = \begin{bmatrix} \boxed{\begin{matrix} \frac{\partial L}{\partial Y_{00}} \\ \frac{\partial L}{\partial Y_{01}} \end{matrix}} & \begin{matrix} \frac{\partial L}{\partial Y_{10}} \\ \frac{\partial L}{\partial Y_{11}} \end{matrix} & \begin{matrix} \frac{\partial L}{\partial Y_{20}} \\ \frac{\partial L}{\partial Y_{21}} \end{matrix} \end{bmatrix}$$

$$\frac{\partial L}{\partial b_0} = \frac{\partial L}{\partial Y_{00}} + \frac{\partial L}{\partial Y_{01}}$$

sum over all samples.