Relu

$$X \in R^{K \times n}$$

K = batch size

n = input dim

$$Y = \text{Relu}(X)$$

$$Y \in R^{K \times n}$$

Forward Propagation

Suppose $X \in R^{1 \times n}$, $Y \in R^{1 \times n}$

$$Y = \text{Relu}(X)$$

$$y_i = \begin{cases} x_i , & x_i \geq 0 \\ 0 , & \text{otherwise} \end{cases} \qquad (1)$$

where $1 \leq i \leq n$

Now for $X \in R^{k \times n}$ $k =$ batch size

$$Y = Relu(X)$$

where $Y \in R^{k \times n}$

$$Y_{ij} = \begin{cases} X_{ij} & , \text{ if } X_{ij} \geq 0 \\ 0 & , \text{ otherwise} \end{cases}$$

where $1 \leq i \leq k$
$1 \leq j \leq n$

Derivative

Suppose $X \in R^{1 \times n}$ and $Y \in R^{1 \times n}$

then, $$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \frac{\partial Y}{\partial X}$$

here $\frac{\partial L}{\partial Y} \in R^{1 \times n}$

Now by equation 1

$$Y = \text{Relu}(x)$$

$$Y_i = \begin{cases} x_i, & x_i \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

then

$$\left(\frac{dY_i}{dx}\right)_j = \begin{cases} 1, & \text{if } i = j \text{ and } x_i > 0 \\ 0, & \text{otherwise} \end{cases}$$

Here $1 \leq i \leq n$
$1 \leq j \leq n$ ——— 2

Now,

$$\frac{dL}{dx} = \frac{dL}{dY} \frac{dY}{dx}$$

then,

$$\left(\frac{\partial L}{\partial x}\right)_i = \begin{cases} \left(\frac{\partial L}{\partial y}\right)_i , & \text{if } \frac{\partial y}{\partial x} = 1 \\ 0 , & \text{otherwise} \end{cases}$$

$$1 \leq i \leq n \quad \text{———} \quad (3)$$

And by equation (2), $\left(\frac{\partial y}{\partial x}\right)_i = 1$ when

$x_i \geq 0$

Now suppose $X \in R^{k \times n}$, $Y \in R^{k \times n}$

then,

We can apply equation (3) for different k input independently

So equation (3) can generalize by

$$\left(\frac{\partial L}{\partial x}\right)_{i,j} = \begin{cases} \left(\frac{\partial L}{\partial y}\right)_{i,j} , & \text{if } x_{ij} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $1 \leq i \leq K$
$1 \leq j \leq n$

$\Rightarrow$ So we can consider $\frac{dY}{dX}$ for each K create a $n \times n$ matrix, so there will K, $n \times n$ matrix. where each matrix can have 1 at diagonal and rest will be zero

To find $\frac{dL}{dX}$ we multiply $\frac{dL}{dY} \in R^{K \times n}$ and $\frac{dY}{dX} \in R^{K \times n \times n}$, Such that each row vector in $\frac{dL}{dY}$ will multiply respective matrix in $\frac{dY}{dX}$.