Linear Layer

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

m = batch size
n = input dim

In some places
k = batch size
also used
and m = output dim

$$Y = Wx + b$$

$$W \in R^{k \times n}, \quad b \in R^k$$

k = output dim
n = Input dim

Forward Propagation

$$Y = Wx^T + b$$

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

$$W \in R^{k \times n}$$

$$b \in R^k$$

$$Y \in R^{k \times m}$$

we will take transpose of y to get $R^{m \times k}$

Suppose $m=1$ on batch size is 1

then $x \in R^{1 \times n}$

$$Y = wx^T + b$$

$$y_i = \sum_{j=1}^{n} w_{ij} x_j + b_i \quad \text{---------- (1)}$$

Now suppose for $m > 1$ on batch size is more than 1

$$y_{ij} = \sum_{d=1}^{n} w_{id} x_{dj} + b_i \quad \text{---------- (2)}$$

⇒ For more than one input, we can repeat equation (1), same thing we do to find derivative.

Derivative

$$Y = wx^T + b$$

$w \in R^{m \times n}$

$x \in R^n$

$b \in R^m$

$$\frac{dL}{dw}, \frac{dL}{db}, \frac{dL}{dx}$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial w} \qquad \frac{\partial L}{\partial b} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial b}$$

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial x}$$

let's Consider for only one case means $x \in R^{1 \times n}$

Previous grad on $\frac{\partial L}{\partial y} \in R^{m \times 1}$ (for case $\frac{\partial L}{\partial y} \in R^{m \times n}$)

then

$$y = wx^T + b$$

by (1)

$$y_i = \sum_{j=1}^{n} w_{ij} x_j + b_i$$

$$\frac{dy_i}{dw} = \frac{d}{dw}\left(\sum_{j=1}^{n} w_{ij} x_j + b_i\right)$$

$$\Rightarrow \frac{d}{dw}\sum_{j=1}^{n} w_{ij} x_j + \frac{d}{dw} b_i$$

$$\Rightarrow \frac{d}{dw}\sum_{j=1}^{n} w_{ij} x_j$$

$$\left(\frac{dy_i}{dw}\right)_{kl} = \begin{cases} x_l & , \text{if } k = i \\ 0 & , \text{otherwise} \end{cases}$$

(3)

$1 \le k \le m$
$1 \le l \le n$

$$\Rightarrow \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & & \\ w_{m1} & \cdots & w_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$\Rightarrow \frac{d}{dw}\begin{bmatrix} w_{11}x_1 + \cdots + w_{1n}x_n \\ \vdots \\ w_{m1}x_1 + \cdots + w_{mn}x_n \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial}{\partial w_{11}}(w_{11}x_1 + \cdots + w_{1n}x_n) & \cdots & \frac{\partial}{\partial w_{m}}(w_{11}x_1 + \cdots + w_{mn}x_n) \\ \vdots & & \\ \frac{\partial}{\partial w_{m1}}(w_{11}x_1 + \cdots + w_{1n}x_n) & \cdots & \frac{\partial}{\partial w_{mn}}(w_{11}x_1 + \cdots + w_{mn}x_n) \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} x_1 & \cdots & x_n \\ & \bigcirc & \end{bmatrix} \quad \text{——— (4)}$$

Now by equation (3) and (4) we can show full derivative

$$\frac{dY}{dw} = \begin{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \\ & \bigcirc & \end{bmatrix} & \cdots & \begin{bmatrix} & \bigcirc & \\ x_1 & \cdots & x_n \end{bmatrix} \end{bmatrix}$$

$$\frac{dY}{dw} \in R^{m \times m \times n} \quad \text{——— (5)}$$

Prev grad, which is gradient of next layer and it will be used in chain rule of derivative to find the grad of loss wrt to W then, to equation (5) we can write

$$\Rightarrow \text{grad} \circledcirc \frac{dy}{dw}$$

here grad is not directly $\frac{dL}{dy}$, because it will be scalar

here,

$\circledcirc \Rightarrow$ matrix multiplication

$$\text{grad} \in R^{1 \times m}$$

$$\boxed{\text{grad} \circledcirc \frac{dy}{dw} = \text{grad}^T \circledcirc x}$$ ——— (6)

Suppose

$$\text{grad} = [g_1 - - - - g_m]$$

$$x = [x_1 - - - - - x_n]$$

then equation (6) will be

$$grad @ \frac{dY}{dw} = \begin{bmatrix} g_1 \\ \vdots \\ g_m \end{bmatrix} [x_1 - - - - x_n]$$

(7)

⇒ Now for batch size K the input x dim will be $X \in R^{K \times n}$, and $grad \in R^{K \times m}$. We have to find derivative by taking mean of every

here, $grad = \frac{dL}{dY} \in R^{K \times m}$

$$\frac{dL}{dw} = \frac{1}{K} \sum_{i=1}^{K} grad_{(i)}^{T} @ x$$

$$\frac{dL}{dw} = \frac{1}{K} \left(grad^{T} @ x \right)$$

$$X \in R^{1 \times n}$$

$$grad \in R^{1 \times m}$$

$$\frac{dL}{dW} \in R^{m \times n}$$

Now let's find $\frac{dL}{db}$, suppose

$Y \in R^{1 \times m}$, $X \in R^{1 \times n}$, $W \in R^{m \times n}$, $b \in R^{1 \times m}$

$$Y = WX + b$$

By equation (1)

$$Y_i = \sum_{j=1}^{n} W_{ij} X_j + b_i$$

$$\frac{dY_i}{db} = \frac{d}{db}\left(\sum_{j=1}^{n} W_{ij} X_j + b_i\right)$$

$$\frac{dY_i}{db} = \frac{d}{db} \sum_{j=1}^{n} W_{ij} X_j + \frac{d}{db} b_i$$

$$\frac{dY_i}{db} = \frac{d}{db} b_i$$

$$\left(\frac{dY_i}{db}\right)_j = \begin{cases} 1 \text{ , if } i = j \\ 0 \text{ , else} \end{cases}$$

$$1 \leq i \leq m$$
$$1 \leq j \leq m$$

——— (8)

To understand equation 8 suppose

$$b = [b_1 \text{ ---- } b_m]$$

$$Y = [Y_1 \text{ ---- } Y_m]$$

$$\frac{dL}{dY} = [g_1 \text{ ---- } g_m]$$

then,

$$\frac{dY}{db} = \begin{bmatrix} \frac{dY_1}{db_1} & \cdots & \frac{dY_m}{db_m} \\ \vdots & & \\ \frac{dY_1}{db_m} & \cdots & \frac{dY_m}{db_m} \end{bmatrix}$$

By equation (8)

$$\boxed{\frac{dY}{db} = I_{m \times m}}$$

——— (9)

So,

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

$$\frac{dL}{db} = \frac{dL}{dY} I$$

$$\boxed{\frac{dL}{db} = \frac{dL}{dY}}$$ ——— (10)

Now for case when batch size = K

then,

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

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

$$\frac{dL}{dY} \in R^{K \times m}$$

we have to find equation (9) for each b_i, $1 \leq i \leq K$,

Now $Y \in R^{K \times m}$ then,

$$\left(\frac{dY_i}{db}\right)_{a,b} = \begin{cases} 1 \text{ , if } a = b \\ 0 \text{ , otherwise} \end{cases}$$

$$1 \leq a \leq m$$
$$1 \leq b \leq m$$
$$1 \leq i \leq K$$

By equation (11), we can see we will get $k, m \times n$ Identity matrix.

To find

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

where,

$$\frac{dL}{dY} \in R^{k \times m}$$

$$\frac{dL}{db} \in R^{k \times m \times m}$$

(By eq (11))

Now each k vector in $\frac{dL}{dY}$ will multiply with respective matrix in $\frac{dL}{db}$ (eq (11)), in return we will get $k \times m$ matrix, $\left(\frac{dY}{db}\right)_i = I$,

$1 \le i \le K$ (By eq (11)) then

Prev grad

$$\frac{dL}{db} = \frac{1}{K} \sum_{i=1}^{K} \left(\frac{dL}{dy}\right)_i$$

Note:

=) when we are finding gradient for batch, Consider each input in batch independent with each other. So while we pass grad to each layer, we can find different grad for different input in batch, and Combine once we find for each input in batch by taking there avg or sum.

=) But instead of finding different

gradint and then take theme avg.
we generaly Combine this process
by taking advantage of algebra
and make single matrix operation

=> so just see diffement independent
(collection of rows / input in batch,
instead of matrix.

Now let's find $\frac{dL}{dx}$, Suppose

$X \in R^{1 \times n}$, $Y \in R^{1 \times m}$, $b \in R^{1 \times m}$, $\frac{dL}{dx} \in R$

$$Y = wx + b$$

By equation (1)

$$Y_i = \sum_{j=1}^{n} w_{ij} x_j + b_i$$

then

$$\frac{dY_i}{dx} = \frac{d}{dx}\left(\sum_{j=1}^{n} w_{ij} x_j + b_i\right)$$

$$\frac{dY_i}{dx} = \frac{d}{dx}\sum_{j=1}^{n} w_{ij} x_j + \frac{d}{dx} b_i$$

$$\frac{dY_i}{dx} = \frac{d}{dx}\sum_{j=1}^{n} w_{ij} x_j$$

$$\frac{dY_i}{dx_j} = W_{ij} \Rightarrow \boxed{\frac{dY}{dx} = W}$$

$$\frac{dY}{dx} \in R^{m \times n}$$

Now to find the $\frac{dL}{dx} = \frac{dL}{dY} @ \frac{dY}{dx}$

Suppose

$$\frac{dL}{dY} = [g_1 \; \cdots \; g_m]$$

$$\frac{dY}{dx} = w = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & & \\ w_{m1} & \cdots & w_{mn} \end{bmatrix}$$

$$\frac{dL}{dx} = [g_1 \cdots g_n] \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & & \\ w_{m1} & \cdots & w_{mn} \end{bmatrix}$$

$$\boxed{\frac{dL}{dx} = [g_1 w_{11} + \cdots + g_m w_{m1}, \; \cdots, \; g_1 w_{1n} + \cdots + g_n w_{mn}]}$$

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

$b \in R^{1 \times m}$, $W \in R^{m \times n}$, $\frac{dL}{dY} \in R^{k \times m}$

$\frac{dY}{dX} \in R^{k \times m}$

Now we have to find derivative for each k batch, this gradient we will provide previous layer.

By equation (12) we can generalize for k batch

$$\frac{dL}{dX} = \frac{dL}{dY} @ \frac{dY}{dX}$$

where, $\frac{dL}{dX} \in R^{k \times n}$, $\frac{dL}{dY} \in R^{k \times m}$

$\frac{dY}{dX} \in R^{m \times n}$