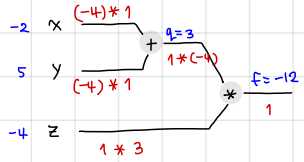
Backpropagation and Neural Networks

1. Backpropagation



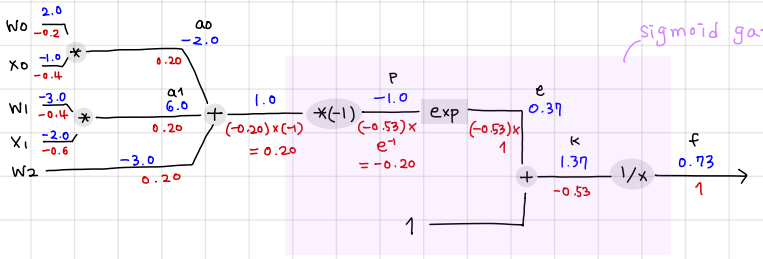
$f(x,y,z) = (x+y)z$

$\begin{cases} q = x+y \\ f = qz \end{cases} \Rightarrow \frac{\partial f}{\partial z} = q = x+y = 3$

$\frac{\partial f}{\partial q} = z \rightarrow \frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial x} = z \cdot 1 = z = -4$

$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial y} = z \cdot 1 = z = -4$

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0 + w_1x_1 + w_2)}}$$



Sigmoid gate : $\sigma(x) = \frac{1}{1+e^{-x}} \rightarrow \frac{\partial \sigma(x)}{\partial x} = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x) \cdot (1-\sigma(x))$

$\rightarrow 0.73 \cdot (1-0.73) = 0.20$

$f = \frac{1}{K} \rightarrow \frac{\partial f}{\partial K} = -\frac{1}{K^2} \rightarrow -\frac{1}{1.37^2} = -0.53$

$K = e+1 \rightarrow \frac{\partial K}{\partial e} = 1 \rightarrow \frac{\partial f}{\partial e} = -0.53$

$e = e^P \rightarrow \frac{\partial e}{\partial P} = e^P \rightarrow \frac{\partial e}{\partial P} = e^{-1} \otimes \frac{\partial f}{\partial e} = -0.53 \rightarrow -0.20$

$a_0 = w_0 \cdot x_0 \rightarrow \frac{\partial a_0}{\partial w_0} = x_0 = -1.0$

$\frac{\partial a_0}{\partial x_0} = w_0 = 2.0$

$a_1 = w_1 \cdot x_1 \rightarrow \frac{\partial a_1}{\partial w_1} = x_1 = -2.0$

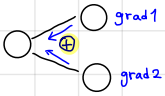
$\frac{\partial a_1}{\partial x_1} = a_1 = -3.0$

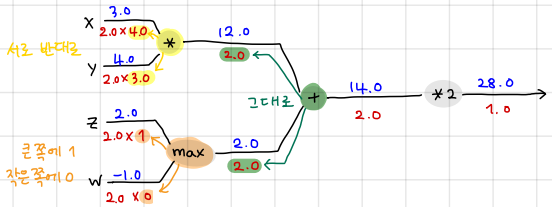
① add gate : gradient distributor

② max gate : gradient router

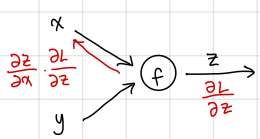
③ mul gate : gradient switcher

④



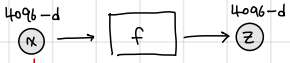


2. Jacobian matrix



$\frac{\partial z}{\partial x}$: Jacobian matrix ; derivatire of each element of z w.r.t each element of x

ex)

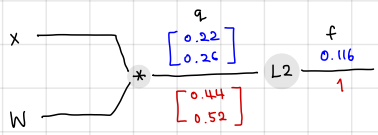


Jacobian $\left(\frac{\partial z}{\partial x}\right)$. shape : 4096 x 4096 $\rightarrow \frac{\partial L}{\partial x} = \frac{\partial z}{\partial x} \cdot \frac{\partial L}{\partial z} = [4096 \times 4096] \cdot [4096 \times 1]$

$= [4096 \times 1]$

3. matmul gate

$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}$



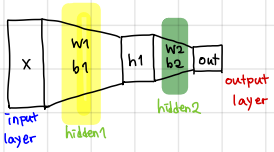
$\begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$

$f = \|q\|^2 \rightarrow \frac{\partial f}{\partial q} = 2q$

$q = xW \rightarrow \frac{\partial q}{\partial x} = W \rightarrow \frac{\partial f}{\partial x} = 2qW^T$

$\rightarrow \frac{\partial q}{\partial W} = x \rightarrow \frac{\partial f}{\partial W} = 2x^T q$

2 Neural Networks : linear layer 여러 개 쌓아서 neural network 구성

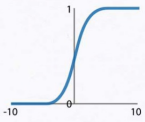


⇒ 2-hidden layer neural net
3-layer neural net

3. Activation Function

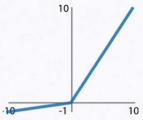
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



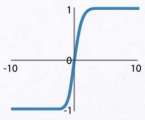
Leaky ReLU

$$\max(0.1x, x)$$



tanh

$$\tanh(x)$$

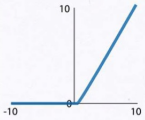


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ReLU

$$\max(0, x)$$



ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$