Forward Pass :-

Backward Pass

* Once we get the output we travel backwards & update the weights

⇒ Derivation of Tan-H activation function

$$g(x) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$\frac{dg(x)}{dx} = \frac{(e^z + e^{-z})\frac{d(e^z - e^{-z})}{dz} - (e^z - e^{-z})\frac{d(e^z + e^{-z})}{dz}}{(e^z + e^{-z})^2}$$

$$= \frac{(e^z + e^{-z})(e^z + e^{-z}) - (e^z - e^{-z})(e^z - e^{-z})}{(e^z + e^{-z})^2}$$

$$= \frac{(e^z + e^{-z})^2 - (e^z - e^{-z})^2}{(e^z + e^{-z})^2}$$

$$= 1 - \frac{(e^z - e^{-z})^2}{(e^z + e^{-z})^2}$$

$$\boxed{g'(x) = 1 - (g(x))^2}$$

⇒ Derivation of Relu Activation Function :-

1) for $z < 0$

$$g(z) = 0$$

$$\boxed{\frac{d(g(z))}{dx} = 0}$$

2) for $z > 0$

$$g(z) = x$$

$$\boxed{\frac{d(g(z))}{dx} = \frac{d(x)}{dx} = 1}$$

⇒ Derivation of Leaky Relu :-

1) for $z < 0$

$$g(z) = \alpha x$$

$$\frac{d(g(z))}{dx} = \frac{d(\alpha x)}{dx}$$

$$= \alpha \frac{d(x)}{dx}$$

$$= \alpha \times 1$$

$$\boxed{g'(x) = \alpha}$$

2) for $z > 0$

$$g(z) = x$$

$$\frac{d(g(z))}{dx} = \frac{d(x)}{dx} \Rightarrow \boxed{g'(z) = 1}$$

network into probabilities.

* These probabilities are distributed across different classes such that their sum equals 1.
* It is suitable for classification tasks

Formula :-

$$S_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

Example :-

$$x = \begin{bmatrix} 2.0 \\ 1.0 \\ 0.1 \end{bmatrix} \xrightarrow{\text{Softmax}} S = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix}$$

$$S_1 = \frac{e^{x_1}}{e^{x_1} + e^{x_2} + e^{x_3}} = \frac{e^2}{e^2 + e^1 + e^{0.1}} = \frac{7.389}{11.21} = 0.66$$

$$S_2 = \frac{e^{x_2}}{e^{x_1} + e^{x_2} + e^{x_3}} = \frac{e^1}{e^2 + e^1 + e^{0.1}} = \frac{2.71}{11.21} = 0.24$$

$$S_3 = \frac{e^{x_3}}{e^{x_1} + e^{x_2} + e^{x_3}} = \frac{e^{0.1}}{e^2 + e^1 + e^{0.1}} = \frac{1.105}{11.21} = 0.10$$

$$\therefore S = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix} = \begin{bmatrix} 0.66 \\ 0.24 \\ 0.10 \end{bmatrix}$$

⇒ Derivative of softmax Activation Function :-

Step 1 :- Define the softmax function

$$\sigma_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

Assume $S = \sum_{j=1}^{n} e^{x_j}$

$$\therefore \sigma_i = \frac{e^{x_i}}{S}$$

We want to compute $\frac{\partial \sigma_i}{\partial z_j}$

Step 2 :- Apply quotient rule

$$\sigma_i = \frac{e^{x_i}}{S}$$

$$\frac{\partial \sigma_i}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\frac{e^{x_i}}{S}\right)$$

$$= \frac{\partial}{\partial x_j}\left(e^{x_i} \cdot \frac{1}{S}\right) \quad \text{\# express as a product}$$

① ⟵ $$= \left(\frac{\partial e^{x_i}}{\partial x_j} \cdot \frac{1}{S}\right) + \left(\frac{\partial}{\partial x_j}\left(\frac{1}{S}\right) \cdot e^{x_i}\right) \quad \text{\# use product rule}$$

② ⟵ $$\frac{\partial e^{x_i}}{\partial x_j} = \begin{cases} e^{z_i} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$S = \sum_{j=1}^{n} e^{x_j} \Rightarrow \frac{\partial S}{\partial x_j} = 0 + 0 + \cdots e^{x_j}$$

$$= e^{x_j}$$

③ ⟵ $$\frac{\partial}{\partial x_j}\left(\frac{1}{S}\right) = -\frac{1}{S^2} \cdot \frac{\partial S}{\partial x_j} \quad \text{\# chain rule of differentiation}$$

$$= -\frac{1}{S^2} \cdot e^{x_j}$$

$$= -\frac{e^{x_j}}{S^2}$$

step 3 : Plug-in ② and ③ in ①

$$\frac{\partial \sigma_i}{\partial x_j} = \left(\frac{\partial}{\partial x_j}(e^{x_i}) \cdot \frac{1}{S}\right) + \left(\frac{\partial}{\partial x_j}\left(\frac{1}{S}\right) \cdot e^{x_i}\right)$$

$$= \left(\frac{\partial}{\partial x_j}(e^{x_i}) \cdot \frac{1}{S}\right) + \left(e^{x_i} \cdot -\frac{e^{x_j}}{S^2}\right)$$

step 4 : Now take seperate cases

case - 1 : $i = j$

$$\frac{\partial(\sigma_i)}{\partial x_j} = \frac{e^{x_i}}{S} - \frac{e^{x_i} e^{x_j}}{S^2}$$

$$= \frac{e^{x_i}}{S} - \frac{e^{x_i} e^{x_i}}{S^2}$$

$$= \sigma_i - \sigma_i^2$$

$$\boxed{\frac{\partial(\sigma_i)}{\partial x_j} = \sigma_i(1 - \sigma_i)}$$

case - 2 : $i \neq j$

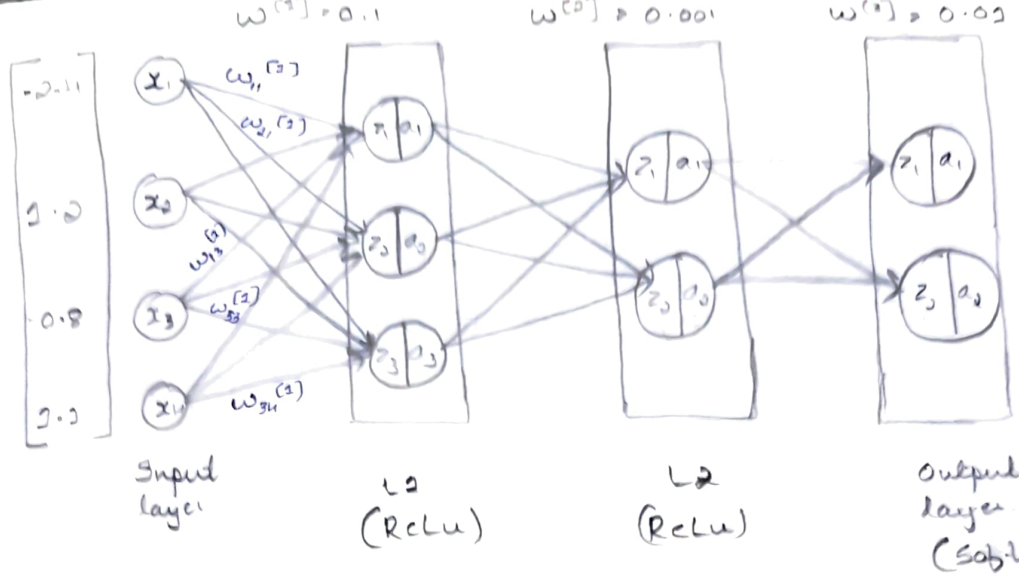
$$\frac{\partial(\sigma_i)}{\partial x_j} = \left(0 \cdot \frac{1}{S}\right) + \left(e^{x_i} \cdot -\frac{e^{x_j}}{S^2}\right)$$

$$= -\frac{e^{x_i} \cdot e^{x_j}}{S^2}$$

$$= -\frac{e^{x_i}}{S} \cdot \frac{e^{x_j}}{S}$$

$$= -S_i \cdot S_j$$

$$\boxed{\frac{\partial(\sigma_i)}{\partial x_j} = -\sigma_i \cdot \sigma_j}$$



⇒ Layer 1

$$z_1 = (-2.4 \times 0.1) + (1.2 \times 0.1) + (-0.8 \times 0.1) + (1.1 \times 0.1)$$

$$= -0.24 + 0.12 - 0.08 + 0.11$$

$$= -0.09$$

$$z_2 = \begin{bmatrix} -2.4 \\ 1.2 \\ -0.8 \\ 1.1 \end{bmatrix} \times \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix} = -0.09$$

$$z_3 = \begin{bmatrix} -2.4 \\ 1.2 \\ -0.8 \\ 1.1 \end{bmatrix} \times \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix} = -0.09$$

After Relu activation all values in layer 1 will become 0

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

⇒ Layer 2 : L_1

$$z_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 0.001 & 0.001 & 0.001 \end{bmatrix} = 0$$

$$z_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 0.001 & 0.001 & 0.001 \end{bmatrix} = 0$$

After Relu activation all values in layer 2 will also will become 0

⇒ Layer 3 [output layer]

$$z_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 0.01 & 0.01 \end{bmatrix} = 0$$

$$z_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 0.01 & 0.01 \end{bmatrix} = 0$$

$$a_1 = \frac{e^0}{e^0 + e^0} = 0.5 \qquad a_2 = \frac{e^0}{e^0 + e^0} = \frac{e^0}{2e^0} = 0.5$$

After softmax activation all values in the output will become 0.5

$$\text{Output Layer} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

⇒ Sizing :-

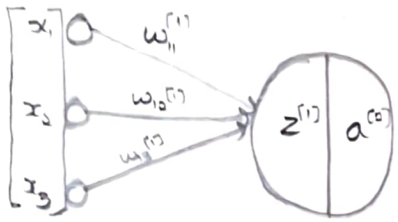
* More the hidden neurons in a layer, more the model will fit to the data

* Due to the presence of multiple deep layer, what we optimize is not a convex function

⇒ Regularization

* overfitting in a neural network can be controlled by applying regularization

⇒ Perceptron :-



$$z^{[1]} = w_{11}^{[1]} x_1 + w_{12}^{[1]} x_2 + w_{13}^{[1]} x_3 + b_1^{[1]}$$

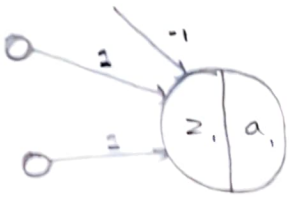
$$a^{[1]} = g(z^1)$$

$$g = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z \leq 0 \end{cases}$$

* Single layer Neural Network with one Neuron is called as a perceptron

⇒ AND

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1



$$z_1 = w_{11}^{[1]} x_1 + w_{12}^{[1]} x + b$$

$$a = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z \leq 0 \end{cases}$$

a) = $0 \times 1 + 0 \times 1 + -1 \Rightarrow -1$ $\quad a = 0 \ (\because z \leq 0)$

b) = $0 \times 1 + 1 \times 1 + -1 = 0$ $\quad a = 0 \ (\because z \leq 0)$

c) = " " " = 0 $\quad a = 0 \ (\because z \leq 0)$

d) = $1 \times 1 + 1 \times 1 + -1 = 1$ $\quad a = 1 \ (\because z > 0)$

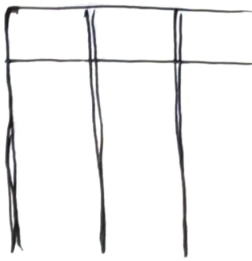
⇒ OR

a) = 0×1 + 0×1 + -1 = -1 → a = 0 (∵ z ≤ 0)

b) = 0 → a = 0 (∵ z ≤ 0)

c) = 0 → a = 0 (∵ z ≤ 0)

d) = 1 → a = 1 (∵ z > 0)



$y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ X , this is not right for OR gate

change bias to 0

a) = 0×1 + 0×1 + 0 = 0 → a = 0 (∵ z ≤ 0)

b) = 0×1 + 1×1 + 0 = 1 → a = 1 (∵ z > 0)

c) = 1×1 + 0×1 + 0 = 1 → a = 1 (∵ z > 0)

d) = 1×1 + 1×1 + 0 = 2 → a = 1 (∵ z > 0)

⇒ XOR

a) = 0×1 + 0×1 + 0 = 0 → a = 0

b) = 0×1 + 1×1 + 0 = 1 → a = 1

c) = 1×1 + 0×1 + 0 = 1 → a = 1

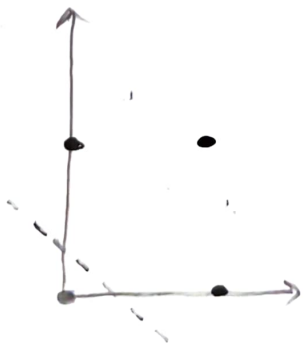
d) = 1×1 + 1×1 + 0 = 2 → a = 1

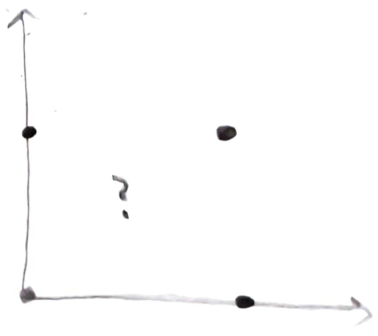
Truth table

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

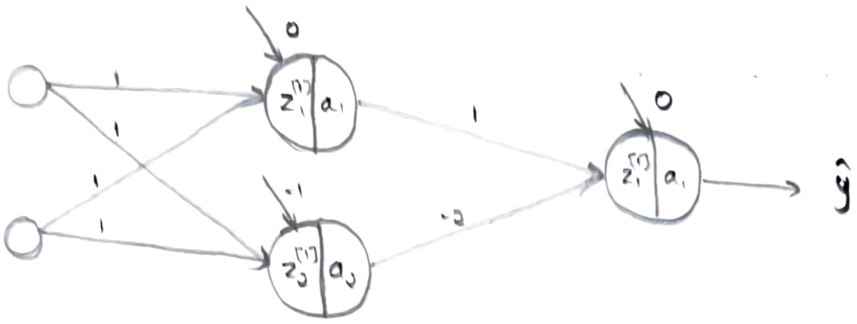
$y = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ X

* A linear boundary cannot be drawn for XOR because XOR is non linear





* A perceptron can model only linear combinations and therefore by introducing an hidden layer we can model non-linear function



$z_1^{[1]} = 0 \times 1 + 0 \times 1 + 0 = 0$ $\quad a = 0 \ (z \leq 0)$

$z^{[1]} = w^{[1]} \times a^{[0]} + b^{[1]}$

$= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix}$

$z^{[1]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$ $\quad a^{[1]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$z^2 = w^{[2]} \times a^{[1]} + b^{[2]}$

$= \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ -2 \end{bmatrix} = 0 + \begin{bmatrix} 1 \\ -2 \end{bmatrix}$

$= \begin{bmatrix} 1 \\ -2 \end{bmatrix} \Rightarrow a = [$

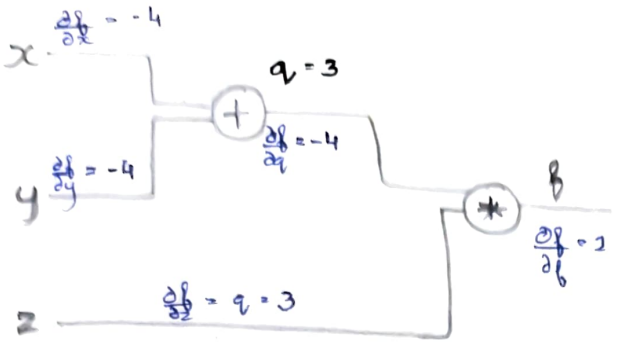
$z^2 = \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0 = 0$

$a^2 = 0 \ (\because z \leq 0)$

$\therefore y = 0$ ($\because$ Row 1 is correct)

⇒ Example

$f = (x+y)z$



Let $q = (x+y)$

$f = qz$

$\frac{\partial q}{\partial x} = 1 \qquad \frac{\partial q}{\partial y} = 1$

$\frac{\partial f}{\partial q} = z \qquad \frac{\partial f}{\partial z} = q$

Computational graph :-

1) $x = -2 \qquad y = 5 \qquad z = 4$ (Random initialization)

$f = (-2+5)4$

$= 12$

$\frac{\partial f}{\partial f} = 1$

$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial x}$ (global gradient · local gradient) $= -4 \times 1 = -4$

chain rule

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

$\frac{\partial f}{\partial z} = q = 3$

$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} * \frac{\partial q}{\partial x}$

$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} * \frac{\partial q}{\partial y}$

Example 2

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

$$x = -2 \quad y = 5 \quad z = -4$$

① Identify additional functions

② Draw a computational graph

③ Perform forward pass

④ Perform backward pass starting from the end of circuit

① Identify additional functions

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

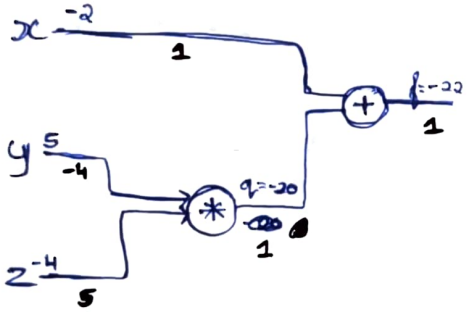
$$q = yz$$

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

$$\frac{\partial f}{\partial x} = 1 \qquad \frac{\partial f}{\partial q} = 1$$

$$\frac{\partial q}{\partial y} = z \qquad \frac{\partial q}{\partial z} = y$$

② Draw a computational graph :-



③ Forward pass :-

$$q = yz = -20$$

$$f = x + q = -22$$

④ Backward pass :-

$$\frac{\partial f}{\partial f} = 1$$

$$\frac{\partial f}{\partial q} = 1 \quad , \quad \frac{\partial f}{\partial x} = 1$$

Global gradient × Local gradient

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

$$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial q} * \frac{\partial q}{\partial z} = 1 * 5 = 5$$

(transformation)

$$f(w,x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$

$w_0 = 2$
$x_0 = -1$
$w_1 = -3$
$x_1 = -2$
$w_2 = -3$

① Identify additional functions

$f_1 = w_0 x_0 \qquad f_2 = w_1 x_1 \qquad f_3 = w_2$

$$f(w,x) = \frac{1}{1 + e^{-(f_1 + f_2 + f_3)}}$$

$f_4 = 1 + e^{-(f_1 + f_2 + f_3)}$

$f(w,x) = \frac{1}{f_4}$

① Identify additional functions :

$f_1 = w_0 x_0 \qquad f_2 = w_1 x_1$

$f_3 = f_1 + f_2 \qquad f_4 = f_3 + w_2$

$f_5 = -f_4$

$f_6 = e^{f_5}$

$f_7 = 1 + f_6$

$f_8 = \frac{1}{f_7}$

$f(w,x) = f_8$

$\frac{\partial f_1}{\partial x_0} = w_0 \quad , \frac{\partial f_2}{\partial x_1} = w_1$

$\frac{\partial f_1}{\partial w_0} = x_0 \quad , \frac{\partial f_2}{\partial w_1} = x_1$

$\frac{\partial f_3}{\partial f_1} = 1 \quad , \frac{\partial f_3}{\partial f_2} = 1$

$\frac{\partial f_4}{\partial f_3} = 1 \quad , \frac{\partial f_4}{\partial w_2} = 1$

$\frac{\partial f_5}{\partial f_4} = -1$

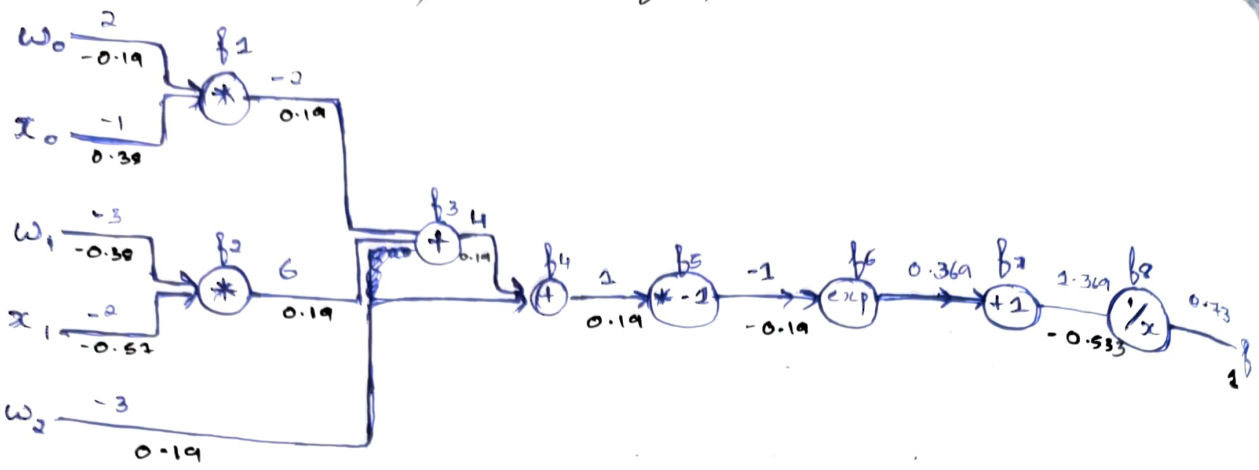
$\frac{\partial f_6}{\partial f_5} = e^{f_5}$

$\frac{\partial f_7}{\partial f_6} = 1$

$\frac{\partial f_8}{\partial f_7} = 0 - \frac{1}{f_7^2}$

$\frac{\partial f}{\partial f_8} = 1$

② Draw a computation graph



③ Forward pass :-

$f_1 = w_0 x_0 = -2$

$f_2 = w_1 x_1 = +6$

$f_3 = -2 + 6 = 4$

$f_4 = 4 - 3 = 1$

$f_5 = 1 * -1 = -1$

$f_6 = e^{-1} = \frac{1}{e} = 0.369$

$f_7 = 1.369$

$f_8 = 0.73$

④ Backward Pass :-

- $\frac{\partial f}{\partial f_8} = 1$

- $\frac{\partial f_8}{\partial f_7} = -\frac{1}{1.369^2} = -0.534$ (local * global) $= \frac{1}{f_7^2} * 1 = -0.533$

- $\frac{\partial f_8}{\partial f_6} =$ local * global $= \partial \cdot * \frac{\partial f_7}{\partial f_6}$

- $\frac{\partial f_8}{\partial f_5} =$ local * global $= \frac{\partial f_6}{\partial f_5} * -0.54 = e^{f_5} * -0.54$

$= e^{-1} * -0.54 = -0.19$

- $\frac{\partial f_6}{\partial f_4} = \frac{\partial f_5}{\partial f_4} * -0.19 = -1 * -0.19 = 0.19$

- $\frac{\partial f_6}{\partial w_2} = \frac{\partial f_4}{\partial w_2} * 0.19 = 1 \times 0.19 = 0.19$

- $\frac{\partial f_6}{\partial f_3} = \frac{\partial f_4}{\partial f_3} * 0.19 = 1 * 0.19 = 0.19$

- $\frac{\partial f_6}{\partial f_2} = \frac{\partial f_3}{\partial f_2} * 0.19 = 1 * 0.19 = 0.19$

- $\frac{\partial f_8}{\partial f_1} = \frac{\partial f_3}{\partial f_1} * 0.19 = 1 \times 0.19 = 0.19$

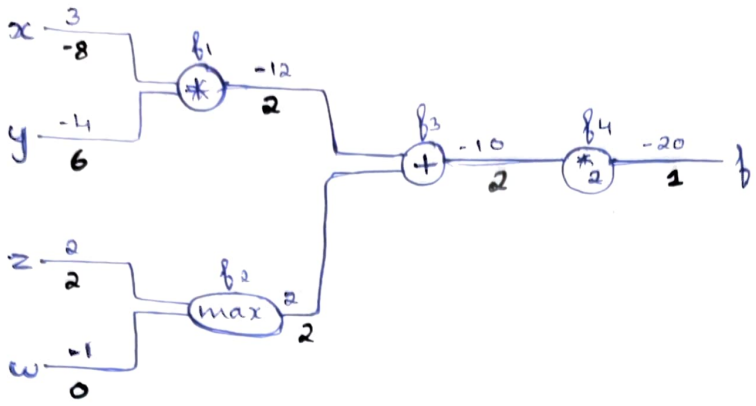
- $\frac{\partial f_6}{\partial x_1} = \frac{\partial f_2}{\partial x_1} * 0.19 = w_1 * 0.19 = -3 \times 0.19 = -0.57$

- $\frac{\partial f_8}{\partial w_1} = \frac{\partial f_2}{\partial w_1} * 0.19 = x_1 * 0.19 = -2 \times 0.19 = -0.38$

- $\frac{\partial f_8}{\partial w_0} = \frac{\partial f_1}{\partial w_0} * 0.19 = -1 * 0.19 = -0.19$

- $\frac{\partial f_8}{\partial x_0} = \frac{\partial f_1}{\partial x_0} * 0.19 = w_0 * 0.19 = 2 \times 0.19 = 0.38$

Example - 4



① Identify additional functions :-

$f_1 = x * y = -12$

$f_2 = \max(z, w) = 2$

$f_3 = f_1 + f_2 = -10$

$f_4 = 2 * f_3 = -20$

$\frac{\partial f_1}{\partial x} = y \qquad \frac{\partial f_1}{\partial y} = x$

$\frac{\partial f_2}{\partial}$

$\frac{\partial f_3}{\partial f_1} = 1 \qquad \frac{\partial f_3}{\partial f_2} = 1$

$\frac{\partial f_4}{\partial f_3} = 2$

④ Backward Pass

- $\frac{\partial f_4}{\partial f_4} = 1$
- $\frac{\partial f_3}{\partial f_4} = \frac{\partial f_4}{\partial f_3} * \text{global} = 2 * 1 = 2$
- $\frac{\partial f_4}{\partial f_1} = \frac{\partial f_3}{\partial f_1} * 2 = 1 * 2 = 2$
- $\frac{\partial f_4}{x} = \frac{\partial f_1}{x} * 2 = y * 2 = -8$
- $\frac{\partial f_4}{y} = \frac{\partial f_1}{y} * 2 = x * 2 = 6$

$f_z = \max(z, w)$

$$\frac{\partial f_z}{\partial z} = \begin{cases} 1 & \text{if } z > w \\ 0 & \text{if } z < w \\ \text{undefined or arbitrary} & z = w \end{cases}$$

$$\frac{\partial f_z}{\partial w} = \begin{cases} 1 & \text{if } w > z \\ 0 & \text{if } w < z \\ \text{undefined or arbitrary} & z = w \end{cases}$$

⇒ Scale and shift

* Mean of the Gaussian will not be 0 and the variance will also be different

⇒ How much to shift ?

* Scale - γ and shift - β and γ and β are learnable parameters

* $\gamma^{[l]}$ $\beta^{[l]}$:- Scale & shift will be applied to each layer

Example :-

$z = [4, 8, 6, 10]$ ⇒ No. of samples in a batch

$\mu = 7$

$$\sigma = \frac{1}{4}\sum_{i=1}^{4}(z_i - \mu)^2 = \frac{(4-7)^2 + (8-7)^2 + (6-7)^2 + (10-7)^2}{4+\epsilon}$$

$\epsilon = 10^{-5}$

$$= \frac{(-3)^2 + 1^2 + 1^2 + 3^2}{4}$$

$\sigma = 5$

$$\tilde{z} = \gamma \hat{z} + \beta$$

$\gamma = 2$

$\beta = 1$

$$\hat{z} = \frac{z - \mu}{\sigma}$$

$\tilde{z}_1 = 2(-1.3416) + 1 = -1.6832$

$\tilde{z}_2 = 2(\hat{z}_2) + 1 = 1.8944$

$\tilde{z}_3 = \ldots$

$\tilde{z}_4 = \ldots$

⇒ Inverted Dropout :-

Train

1) Sample 1 ⟶ output = 2
2) Sample 2 ⟶ dropout ⟶ output = 0
3) Sample 3 ⟶ dropout ⟶ output = 0
4) Sample 4 ⟶ output = 2

Total output = 4

Test

1) Sample 1 ⟶ output = 2
2) Sample 2 ⟶ output = 2
3) Sample 3 ⟶ output = 2
4) Sample 4 ⟶ output = 2

Total output = 8

* To achieve the test output during training we scale the total output @ by p

$$\therefore \text{Total output} = \frac{\text{Total output (train)}}{p}$$

* Since an FNN cannot be applied, we build something different called 'Convolutional Neural Network'.

⇒ Convolution operation :- (1)

* It is the heart of a CNN

3	0	1	2	7	4
1	5	8	9	3	1
2	7	2	5	1	3
0	1	3	1	7	8
4	2	1	6	2	8
2	4	5	2	3	9

Input 6×6 info.

* Computation Operator

1	0	-1
1	0	-1
1	0	-1

filter / kernel
3×3

=

-5	-4	0	8
-10	-2	2	3
0	-2	-4	-7
-3	-2	-3	-16

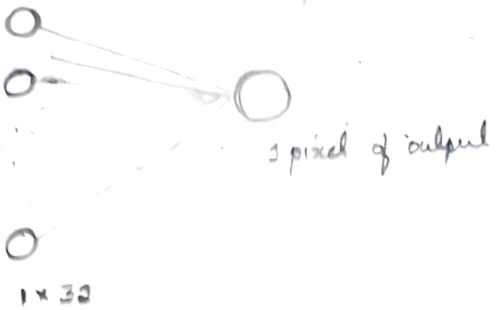
Super-impose filter on the info and then multiply the value (ie perform convolutional operation) of the filter with the info

$= 3 \times 1 + 0 \times 0 + -1 \times 1 + 1 \times 1 + 5 \times 0 + 8 \times -1 + 2 \times 1 + 7 \times 0 + 2 \times -$

$= -5$

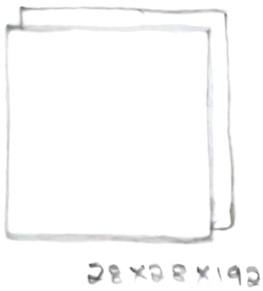
* Now add this value

what it looks like :- for each pixel

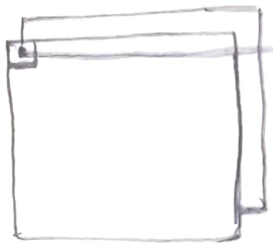


MLP

- Weight sharing : we tell the filter to slide across the whole image and learn what its assigned to.



CONV
5 × 5 × 192 filter
filter = 32
same convolution



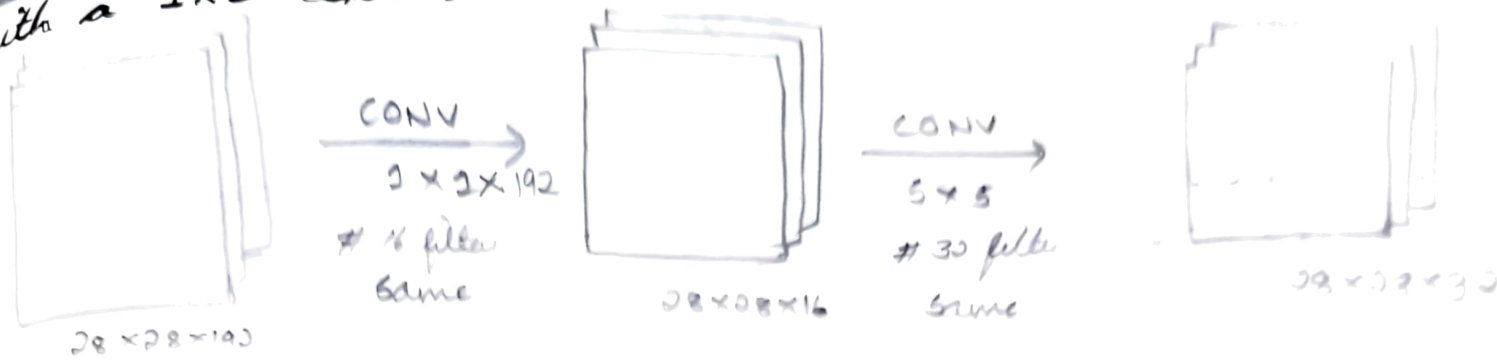
→ To get this one pixel we need to do 5×5×192 operation

* for whole image then;
(5 × 5 × 192) × (28 × 28 × 32)

Total operations = $5 \times 5 \times 192 \times 28 \times 28 \times 32 = 120422400$

- $5 \times 5 \times 192$: filter size
- 28×28 : No. of strides required to cover the whole picture
- 32 : No. of filter

⇒ With a 1×1 CONV :-



Total operations

1) (1×1×192×28×28) × 16 = 2,422,016 ≈ 2 million

$28 = n + 2p - f + 1$

$28 = 28 + 2p - 5 + 1$

$4 = 2p$

$p = 2$

2) 5×5×16×28×28×32 = 10,035,200 ≈ 10 million

Total = 12.5 million

* Almost 1/10th of the total operations as that of model without 1×1 conv.

⇒ Modifying height and width :-

1) Padding.
2) Striding
3) Poding

⇒ Modify channels :-

1) No of filters
2) 1×1 CONV

T = 2 (Two time steps)

$X^{(2)}$: $x_1 \in \mathbb{R}^2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $x_2 \in \mathbb{R}^2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

input dimension = 2 $\Longrightarrow x_t \in \mathbb{R}^2$

hidden state dimensions = 3 $\Longrightarrow h_{(t)} \in \mathbb{R}^3$

$y_t \in \mathbb{R}^2$

$h_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times1}$, $W_{hx} = \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}_{3\times2}$

$W_{hh} = \begin{bmatrix} 0.1 & 0.4 & 0.0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix}_{3\times3}$

$W_{yh} = \begin{bmatrix} 1 & -2 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix}_{2\times3}$

$h_1 = W_{HH} * h_0 + W_{Hx} x_1$

$h_{3\times1} = (3\times3) * (3\times1) + (3\times2) * (2\times1)$

Both result must be (3×1)

$= (3\times1) + (3\times1)$

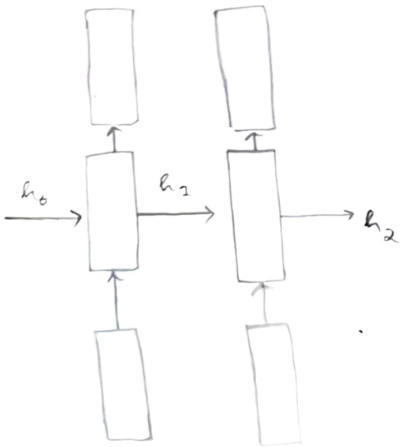
$= (3\times1)$

$y(t) = W_{HY} * h_1$

$y(t)_{2\times1} = W_{HY} * h(t)$

$= (2\times3) * (3\times1)$

$y(t) = (2\times1)$



⇒ Time-step 1 , t = 1

$h_1 = \tanh(W_{HH} * h_0) + (W_{Hx} x_t)$

$W_{HH} * h_0 = \begin{bmatrix} 0.1 & 0.4 & 0.0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix} * \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$W_{Hx} x_t = \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix} * \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$= \begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix}$

$$h_1 = \tanh\left[\begin{bmatrix}0\\0\\0\end{bmatrix} + \begin{bmatrix}-0.1\\1.2\\0.9\end{bmatrix}\right]$$

$$= \tanh\left(\begin{bmatrix}-0.1\\1.2\\0.9\end{bmatrix}\right)$$

$$h_1 = \begin{bmatrix}-0.099\\0.834\\0.71\end{bmatrix}$$

$$y^{(1)} = W_{HY} * h_1$$

$$= \begin{bmatrix}1 & -2 & 0.5\\0.5 & 0.5 & -0.5\end{bmatrix} * \begin{bmatrix}-0.099\\0.834\\0.71\end{bmatrix}$$

$$= \begin{bmatrix}-0.59\\0.01\end{bmatrix}$$

⇒ Time-step 2, t = 2

$$h_2 = \tanh\left(W_{HH} * h_1\right) + \left(W_{HH} * x_2\right)$$

$$W_{HH} * h_1 = \begin{bmatrix}0.1 & 0.4 & 0.0\\-0.2 & 0.3 & 0.2\\0.05 & -0.1 & 0.2\end{bmatrix} * \begin{bmatrix}-0.099\\0.834\\0.71\end{bmatrix}$$

$$= \begin{bmatrix}0.3237\\0.412\\0.0536\end{bmatrix}$$

$$W_{HX} * x_2 = \begin{bmatrix}0.1 & 0.4 & 0.0\\-0.2 & 0.3 & 0.2\end{bmatrix}$$

$$= \begin{bmatrix}0.5 & -0.3\\0.8 & 0.2\\0.1 & 0.4\end{bmatrix} * \begin{bmatrix}-1\\1\end{bmatrix}$$

$$= \begin{bmatrix}-0.8\\-0.6\\0.3\end{bmatrix}$$

$$h_2 = \tanh\left(\begin{bmatrix}0.3237\\0.412\\0.0536\end{bmatrix} + \begin{bmatrix}-0.8\\-0.6\\0.3\end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix}-0.47\\-0.19\\0.3536\end{bmatrix}\right)$$

$$h_2 = \begin{bmatrix}-0.66\\-0.19\\0.337\end{bmatrix}$$

⇒ Example :-

$x_t = [0.5, -0.1]$

$h_{t-1} = [0.0, 0.1]$

$C_{t-1} = [0.2, -0.2]$

$$W_{xi} = \begin{bmatrix} 0.5 & -0.3 \\ 0.4 & 0.1 \end{bmatrix} \qquad W_{hi} = \begin{bmatrix} 0.1 & 0.2 \\ -0.2 & 0.05 \end{bmatrix}$$

$$W_{xf} = \begin{bmatrix} -0.4 & 0.2 \\ 0.3 & 0.3 \end{bmatrix} \qquad W_{hf} = \begin{bmatrix} 0.05 & -0.1 \\ 0.2 & 0.1 \end{bmatrix}$$

$$W_{xo} = \begin{bmatrix} 0.3 & 0.25 \\ -0.2 & 0.1 \end{bmatrix} \qquad W_{ho} = \begin{bmatrix} 0.15 & 0.05 \\ 0.1 & 0.2 \end{bmatrix}$$

$$W_{xg} = \begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix} \qquad W_{hg} = \begin{bmatrix} 0.2 & 0.1 \\ -0.1 & 0.05 \end{bmatrix}$$

$h_t = ?$ $\qquad C_t = ?$

⇒ $f_t = \sigma\left(W_{hf} h_{t-1} + W_{xf} x_t\right)$

$$W_{hf} h_{t-1} = \begin{bmatrix} 0.05 & -0.1 \\ 0.2 & 0.1 \end{bmatrix} * \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}$$

$2 \times 2 \qquad 2 \times 1$

$$= \begin{bmatrix} -0.01 \\ 0.01 \end{bmatrix}_{2 \times 1}$$

$$W_{xf} x_t = \begin{bmatrix} -0.4 & 0.2 \\ 0.3 & 0.3 \end{bmatrix} * \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$$

$$= \begin{bmatrix} -0.22 \\ 0.12 \end{bmatrix}$$

$$\left(W_{hf} h_{t-1} + W_{xf} x_t\right) = \begin{bmatrix} -0.01 \\ 0.01 \end{bmatrix} + \begin{bmatrix} -0.22 \\ 0.12 \end{bmatrix} = \begin{bmatrix} -0.23 \\ 0.13 \end{bmatrix}$$

$$\sigma\left(W_{hf} h_{t-1} + W_{xf} x_t\right) = \sigma \begin{bmatrix} -0.23 \\ 0.13 \end{bmatrix} = \begin{bmatrix} 0.443 \\ 0.532 \end{bmatrix} = f_t \qquad \text{— (1)}$$

$$\Rightarrow i_t = \sigma\left(W_{hi} h_{t-1} + W_{xi} x_t\right)$$

$$W_{hi} h_{t-1} = \begin{bmatrix} 0.1 & 0.2 \\ -0.2 & 0.05 \end{bmatrix} * \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.02 \\ 0.005 \end{bmatrix}$$

$$W_{xi} x_t = \begin{bmatrix} 0.5 & -0.3 \\ 0.4 & 0.2 \end{bmatrix} * \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.28 \\ 0.19 \end{bmatrix}$$

$$\left(W_{hi} h_{t-1} + W_{xi} x_t\right) = \begin{bmatrix} 0.02 \\ 0.005 \end{bmatrix} + \begin{bmatrix} 0.28 \\ 0.19 \end{bmatrix}$$

$$= \begin{bmatrix} 0.30 \\ 0.195 \end{bmatrix}$$

$$\sigma\left(W_{hi} h_{t-1} + W_{xi} x_t\right) = \sigma \begin{bmatrix} 0.30 \\ 0.195 \end{bmatrix} = \begin{bmatrix} 0.575 \\ 0.549 \end{bmatrix} = i_t \longrightarrow (2)$$

$$\Rightarrow o_t = \sigma\left(W_{ho} h_{t-1} + W_{xo} x_t\right)$$

$$W_{ho} h_{t-1} = \begin{bmatrix} 0.15 & 0.05 \\ 0.2 & 0.2 \end{bmatrix} * \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.005 \\ 0.02 \end{bmatrix}$$

$$W_{xo} x_t = \begin{bmatrix} 0.3 & 0.25 \\ -0.2 & 0.2 \end{bmatrix} * \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.125 \\ -0.12 \end{bmatrix}$$

$$\left(W_{ho} h_{t-1} + W_{xo} x_t\right) = \begin{bmatrix} 0.005 \\ 0.02 \end{bmatrix} + \begin{bmatrix} 0.125 \\ -0.12 \end{bmatrix}$$

$$= \begin{bmatrix} 0.130 \\ -0.10 \end{bmatrix}$$

$$\sigma\left(W_{ho} h_{t-1} + W_{xo} x_t\right) = \sigma \begin{bmatrix} 0.130 \\ -0.10 \end{bmatrix}$$

$$= \begin{bmatrix} 0.53 \\ 0.47 \end{bmatrix} = o_t \longrightarrow (3)$$

$\Rightarrow g_t = \tanh(W_{hg} h_{t-1} + W_{xg} x_t)$

$W_{hg} h_{t-1} = \begin{bmatrix} 0.2 & 0.1 \\ -0.1 & 0.05 \end{bmatrix} * \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}$

$= \begin{bmatrix} 0.01 \\ 0.005 \end{bmatrix}$

$W_{xg} x_t = \begin{bmatrix} -0.5 & 0.4 \\ 0.0 & -0.3 \end{bmatrix} * \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$

$= \begin{bmatrix} -0.29 \\ 0.13 \end{bmatrix}$

$(W_{hg} h_{t-1} + W_{xg} x_t) = \begin{bmatrix} 0.01 \\ 0.005 \end{bmatrix} + \begin{bmatrix} -0.29 \\ 0.13 \end{bmatrix}$

$= \begin{bmatrix} -0.28 \\ 0.135 \end{bmatrix}$

$g = (W_{hg} h_{t-1} + W_{xg} x_t) = \tanh \begin{bmatrix} -0.28 \\ 0.135 \end{bmatrix} = \begin{bmatrix} -0.27 \\ 0.134 \end{bmatrix}$ — (4)

$\Rightarrow C_t = f_t \odot C_{t-1} + i_t \odot g_t$

$f_t \odot C_{t-1} = \begin{bmatrix} 0.443 \\ 0.532 \end{bmatrix} \odot \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix} = \begin{bmatrix} 0.0886 \\ -0.1064 \end{bmatrix}$

$i_t \odot g_t = \begin{bmatrix} 0.575 \\ 0.549 \end{bmatrix} \odot \begin{bmatrix} -0.28 \\ 0.135 \end{bmatrix} = \begin{bmatrix} -0.161 \\ 0.074 \end{bmatrix}$

$C_t = \begin{bmatrix} 0.088 \\ -0.106 \end{bmatrix} + \begin{bmatrix} -0.161 \\ 0.074 \end{bmatrix}$

$= \begin{bmatrix} -0.073 \\ -0.03 \end{bmatrix}$

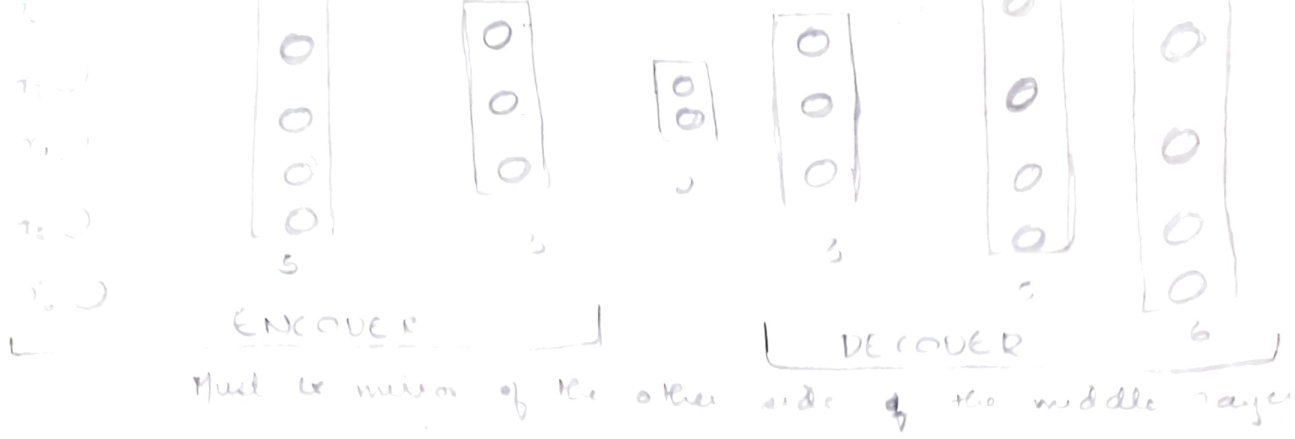
$\Rightarrow h_t = o_t \odot \tanh(c_t)$

$\tanh(c_t) = \tanh \begin{bmatrix} -0.07 \\ -0.03 \end{bmatrix}$

$= \begin{bmatrix} -0.07 \\ -0.03 \end{bmatrix}$

$h_t = \begin{bmatrix} 0.53 \\ 0.47 \end{bmatrix} * \begin{bmatrix} -0.07 \\ -0.03 \end{bmatrix}$

$= \begin{bmatrix} -0.0371 \\ -0.0141 \end{bmatrix}$



Must be mirror of the other side of the middle layer

→ Example :-

$x \in \mathbb{R}^2$, $x = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$

$h \in \mathbb{R}^1$ = ?

w_e = ? = 1×2 = $[0.5, -1.0]$

w_d = ? = 2×1 = $\begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$

$MSE = \frac{1}{2} \| \hat{x} - x \|^2$

$h = w_e x$

$= 2 \times 0.5 + 0 \times -1.0$

$= 1$

decoder = $\begin{bmatrix} 1 \\ 0.5 \end{bmatrix} * 1$ $(w_d * h)$

$= \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$

$L = \frac{1}{2} (1 - \hat{2})^2 + (0.5 - 0.0)^2$ 0.625

$= \frac{1}{2} (1^2) + (0.5)^2$

$= \frac{1 + 0.25}{2}$

$= 0.625$

$h_1 = [1, 0, 1]$

$h_2 = [0, 1, 1]$

$h_3 = [1, 1, 0]$

$s_{t-1} = [1, 0, 1]$ $\qquad c_t = ?$

$$c_t = \sum_{j=1}^{t} \alpha_{t,j} h_j$$

$$= \alpha_{t,1} h_1 + \alpha_{t,2} h_2 + \alpha_{t,3} h_3$$

$$\alpha_{t,1} = \frac{\exp(\text{score}(s_{t-1}, h_1))}{\exp(\text{score}(s_{t-1}, h_1)) + \exp(\text{score}(s_{t-1}, h_2)) + \exp(\text{score}(s_{t-1}, h_3))}$$

$\text{score}(s_{t-1}, h_1) = [1, 0, 1] \cdot [1, 0, 1]$

$= 2$

$\text{score}(s_{t-1}, h_2) = [1\ 0\ 1] \cdot [0, 1, 1]$

$= 1$

$\text{score}(s_{t-1}, h_3) = [1, 0, 1] \cdot [1, 1, 0]$

$= 1$

$$\alpha_{t,1} = \frac{e^2}{e^2 + e^1 + e^1} = \frac{7.4}{7.4 + 2.71 + 2.71} = \frac{7.4}{12.82} = 0.578$$

$$\alpha_{t,2} = \frac{e^1}{e^2 + e^1 + e^1} = \frac{2.71}{12.82} = 0.21$$

$$\alpha_{t,3} = \frac{e^1}{e^2 + e^1 + e^1} = \frac{2.71}{12.82} = 0.21$$

$$c_2 = \alpha_{2,1} h_1 + \alpha_{2,2} h_2 + \alpha_{2,3} h_3$$

$$= 0.58 [1, 0, 1] + 0.2 [0, 1, 1] + 0.2 [1, 2, 0]$$

$$= [0.58, 0, 0.58] + [0, 0.2, 0.2] + [0.2, 0.2, 0]$$

$$= [0.78, 0.4, 0.78]$$

→ # This is more close to h_2, it means, it paying more attention to h_2

Thinking Machines

Thinking	Machines
① q_1 = Thinking k_1 = Thinking	① q_1 = Machines k_1 = Thinking
② q_2 = Thinking k_1 = Machines	q_1 = Machines k_2 = Machines

⇒ Example

Input :- Playing outside ⟹ z_1, z_2

Playing	Outside
$q_1 = [0.212, 0.04, 0.63, 0.36]^T$	$q_2 = [0.1, 0.14, 0.86, 0.77]^T$
$k_1 = [0.31, 0.84, 0.963, 0.57]^T$	$k_2 = [0.45, 0.94, 0.73, 0.58]^T$
$v_1 = [0.36, 0.83, 0.1, 0.38]^T$	$v_2 = [0.31, 0.36, 0.19, 0.72]^T$

Playing	Outside
dot prod. ① $q_1 * k_1 =$ 0.91	① $q_2 * k_2 = 1.25$
② $q_1 * k_2 = 0.80$	⑤ $q_2 * k_1 = 1.41$
⇒ divide by $\sqrt{R^E} = \sqrt{4} = 2$	
① 0.455	① 0.625
② 0.40	② 0.705
⇒ Softmax	
① = 0.51	① 0.48
② = 0.49	② 0.52

⇒ $Softmax_1 * V_1 + Softmax_2 * V_2$

$V_{11} = [0.18, 0.42, 0.05, 0.19]$

$V_{12} = [0.15, 0.18, 0.09, 0.35]$

$0.48 * V_2$

$V_{21} = [0.15, 0.17, 0.09, 0.35]$

$V_{22} = [0.19, 0.43, 0.05, 0.20]$

$0.52 * [V_1]$

⇒ Sum

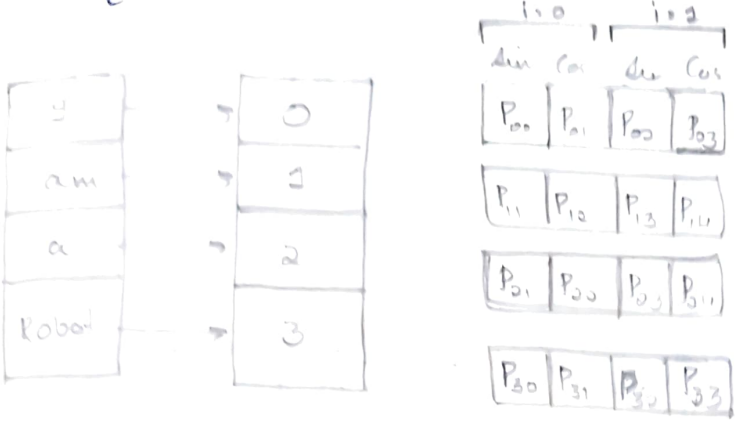
$z_1 = [0.33, 0.6, 0.14, 0.54]$

$z_2 = [0.34, 0.6, 0.14, 0.55]$

Example :

→ I am a Robot

$P_t : d = 4 = \mathbb{R}^4$ # dimension of embedding space



→ $0 \leq i \leq d$

$$P_{Pos,2i} = \sin\left(\frac{pos}{100^{(2i/4)}}\right) \quad \# \mathbb{R}^d = 4$$

$$P_{pos,2i+1} = \cos\left(\frac{pos}{100^{(2i/4)}}\right)$$

⇒ i = 0

* $P_{0,0} = \sin\left(\frac{0}{100^{(0)}}\right) = \sin(0) = 0$

* $P_{0,1} = \cos\left(\frac{0}{100^{0}}\right) = \cos(0) = 1$

⇒ i = 1

* $P_{0,2} = \sin\left(\frac{0}{100^{(0/4)}}\right) = \sin(0) = 0$

* $P_{0,3} = \cos\left(\frac{0}{100^{(0/4)}}\right) = \cos(0) = 1$

$P^{(0)} =$ | 0 | 1 | 0 | 1 |

$\Rightarrow P^1$

$\Rightarrow i = 0$

$P_{1,0} = \sin\left(\frac{1}{100^{(0)}}\right) = \sin(1) = 0.84$

$P_{1,1} = \cos\left(\frac{1}{100^{(0)}}\right) = \cos(1) = 0.54$

$\Rightarrow i = 1$

$P_{1,2} = \sin\left(\frac{1}{100^{0.5}}\right) = \sin\left(\frac{1}{10}\right) = \sin(0.1)$

$= 0.0998$

$= 0.10$

$P_{1,3} = \cos\left(\frac{1}{10}\right) = \cos(0.1) = 0.99 = 1$

$P^1 =$

0.84	0.54	0.1	1

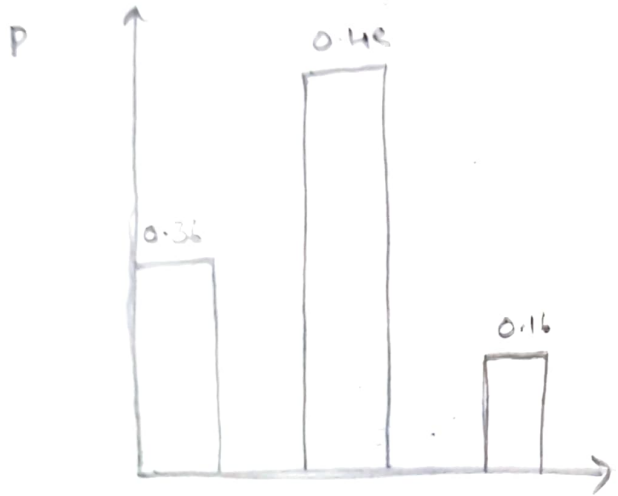
$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \; \mathbb{E} \log\left(\frac{P_x(x)}{P_\theta(x)}\right) = \underset{\theta \in \Theta}{\operatorname{argmin}} \; \mathbb{E} \log(P_x(x)) - \log(P_\theta(x))$$

as we are minimizing w.r.t θ only

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \; \mathbb{E}_{x \sim P_x} -\log(P_\theta(x))$$

Maximal likelihood estimation

⇒ Example :-



Distribution $x \rightarrow$	0	1	2
$P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
$Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$$D_{KL}(P \| Q) = \sum_{x \in X} P(X=x) \ln \frac{P(X=x)}{Q(X=x)}$$

$$= \frac{9}{25} \ln\left(\frac{9/25}{1/3}\right) + \frac{12}{25} \ln\left(\frac{12/25}{1/3}\right) + \frac{4}{25} \ln(\quad)$$

$$= 0.0852$$

$$(QIIP) = \sum_{x \in X} P(X=x) \ln \frac{Q(X=x)}{P(X=x)}$$

$$= \frac{1}{3} \ln \left(\frac{1/3}{9/25} \right) + \frac{1}{3} \ln \left(\frac{1/3}{12/25} \right) + \frac{1}{3} \ln \left(\frac{1/3}{4/25} \right)$$

$$= \frac{1}{3} \ln [0.926] + \frac{1}{3} \ln (0.69) + \frac{1}{3} \ln (2.08)$$

$$= \frac{1}{3} (-0.07) + \frac{1}{3} (-0.37) + \frac{1}{3} \ln (0.73)$$

$$= -0.02 - 0.12 + 0.244$$

$$= 0.10$$