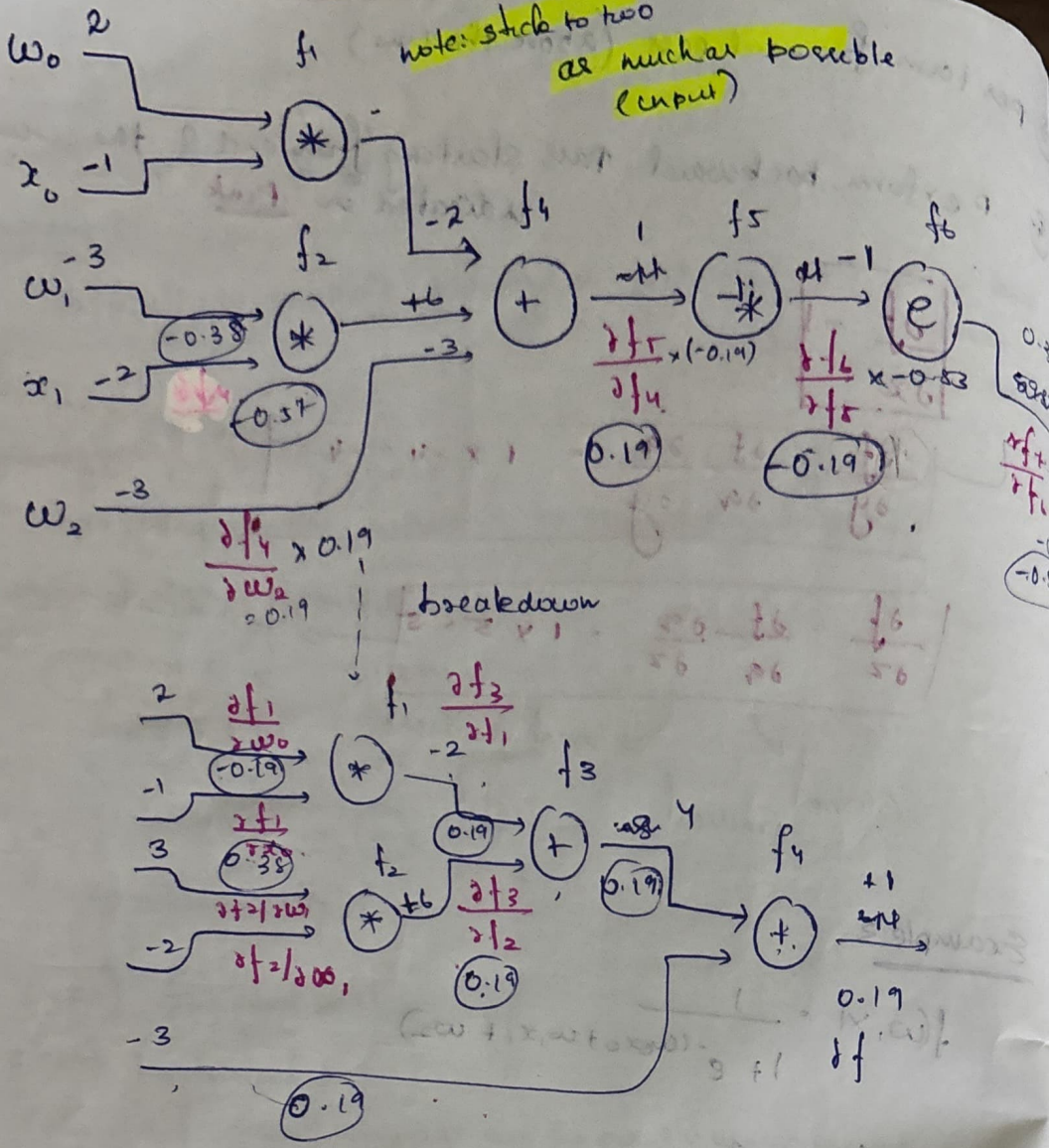


$$\frac{\partial f_1}{\partial w_0} = x_0 \quad (1)$$

$$\frac{\partial f_1}{\partial x_0} = w_0 \quad (2)$$

$$\frac{\partial f_2}{\partial w_1} = x_1 \quad (2)$$

$$\frac{\partial f_2}{\partial x_1} = w_1 \quad (3)$$

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

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

$$\frac{\partial f_4}{\partial f_3} = 1$$

$$\frac{\partial f_4}{\partial w_2} = 1$$

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

$$\frac{\partial f_6}{\partial f_5} = e^{f_5} = 0.37$$

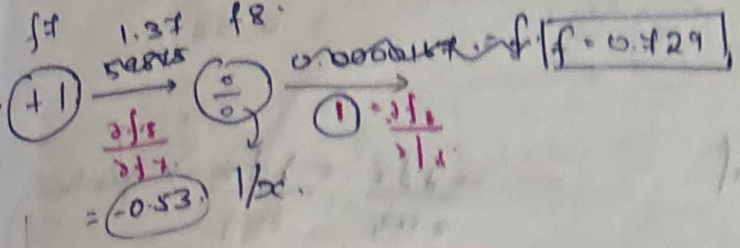
$$\frac{\partial f_7}{\partial f_6} = 1$$

$$\frac{\partial f_8}{\partial f_7} = \frac{-1}{(f_7)^2} = \frac{-1}{(1.37)^2} \approx -0.53$$

$$\frac{\partial f_1}{\partial f_0} = 1$$

$\rightarrow \frac{\partial f_8}{\partial f_8} = 1$ $\quad f_8 = \frac{1}{f_7}$

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



equations

$w_0 x_0 = f_1$

$w_1 x_1 = f_2$

$f_1 + f_2 = f_3$

$f_3 + w_2 = f_4$

$(-1) * f_4 = f_5$

$e^{f_5} = f_6$

$1 + f_6 = f_7$

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

$f_1 = -2$	$f_5 = -1$
$f_2 = 6$	$f_6 = 0.37$
$f_3 = -8.4$	$f_7 = 1.37$
$f_4 = -8.4$	$f_8 = 0.729$

Backward propagation

$\rightarrow \frac{\partial f_8}{\partial f_7} = \text{local} * \text{global}$

$= \frac{-1}{(f_7)^2} * 1$

$= -0.533.$

$\rightarrow \frac{\partial f_8}{\partial f_6} = \text{local} \times \text{global}$

$= \frac{\partial f_7}{\partial f_6} \times (-0.53)$

$= 1 \times (-0.53)$

$= -0.53$

$$\rightarrow \frac{\partial f_8}{\partial f_5} = \frac{\partial f_6}{\partial f_5} \times \text{global}$$

$$= 0.37 \times (-0.53)$$

$$= -0.19$$

$$\rightarrow \frac{\partial f_8}{\partial x_0} = \frac{\partial f_0}{\partial x_0} \times \text{global}$$

$$= w\ 2 \times 0.19$$

$$= 0.38$$

$$\rightarrow \frac{\partial f_8}{\partial f_4} = \frac{\partial f_5}{\partial f_4} \times \text{global}$$

$$= (-1) \times -0.19$$

$$= 0.19$$

$$\rightarrow \frac{\partial f_8}{\partial w_1} = \frac{\partial f_2}{\partial w_1} \times \text{global}$$

$$= -2 \times 0.19$$

$$= -0.38$$

$$\rightarrow \frac{\partial f_8}{\partial w_2} = \frac{\partial f_4}{\partial w_2} \times \text{global}$$

$$= 0.19$$

$$\rightarrow \frac{\partial f_8}{\partial x_1} = \frac{\partial f_2}{\partial x_1} \times \text{global}$$

$$= -3 \times 0.19$$

$$= -0.57$$

$$\rightarrow \frac{\partial f_8}{\partial f_3} = \frac{\partial f_4}{\partial f_1} \times \text{global}$$

$$= 1 \times 0.19$$

$$= 0.19$$

$$\rightarrow \frac{\partial f_8}{\partial f_2} = \frac{\partial f_3}{\partial f_2} \times \text{global}$$

$$= 1 \times 0.19$$

$$= 0.19$$

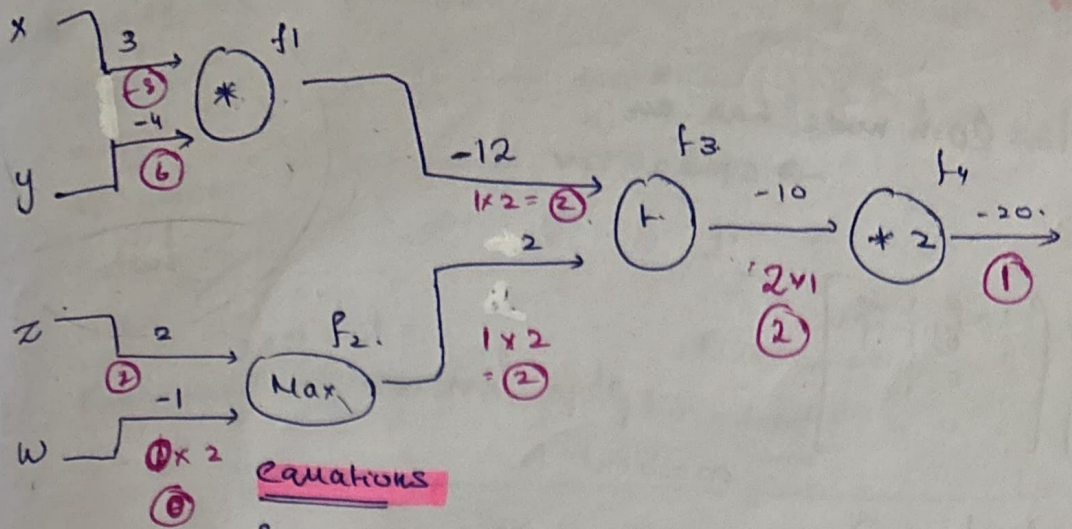
$$\rightarrow \frac{\partial f_8}{\partial f_1} = \frac{\partial f_3}{\partial f_4} \times \text{global}$$

$$= 0.19$$

$$\rightarrow \frac{\partial f_8}{\partial w_0} = \frac{\partial f_1}{\partial w_0} \times \text{global}$$

$$= (1) \times 0.19$$

$$= -0.19$$



equations

$f_1 = x * y$

$f_2 = \max(z, w)$

$f_3 = f_1 + f_2$.

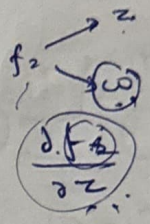
$f_4 = f_3 * 2$.

$x = 3$

$y = -4$

$z = 2$.

$w = -1$.



$\frac{\partial f_1}{\partial x} = y$; $\frac{\partial f_1}{\partial y} = x = 3$.

$\frac{\partial f_2}{\partial z} = 1$; $\frac{\partial f_2}{\partial w} = 0$

$f_2 = \begin{cases} z & z > w \\ w & z < w \end{cases}$

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

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

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

$\partial f_2 / \partial z$.

$f_2' \begin{cases} 1 & z > w \\ 0 & z < w \end{cases}$

undefined $z = w$.

arbitrarily choose 1

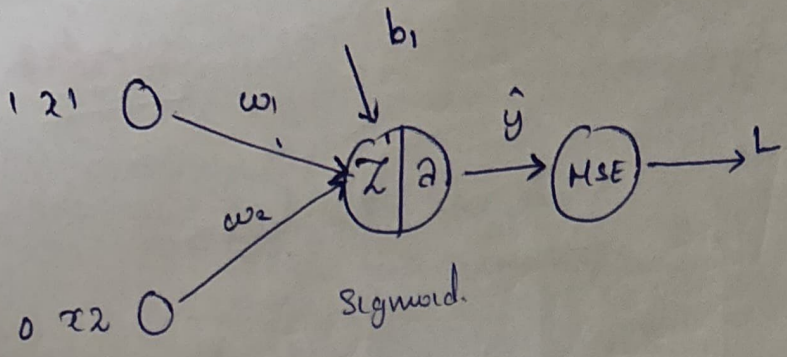
Since $z > w$

$\frac{\partial f_2}{\partial z} = 1$

$\frac{\partial f_2}{\partial w} = 0$.

$\partial f_2 / \partial w$.

$f_2' \begin{cases} 0 & z > w \\ 1 & z < w \end{cases}$

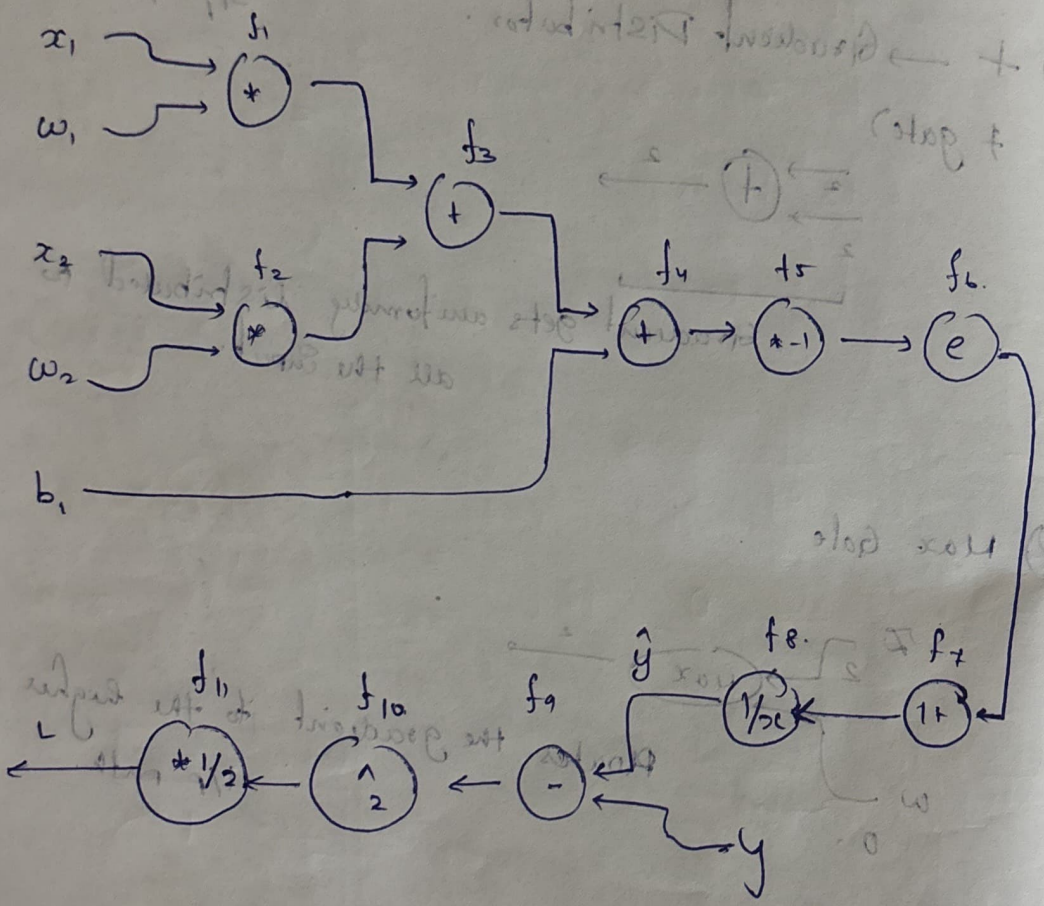


MSE = $\frac{1}{2}(\hat{y}-y)^2$

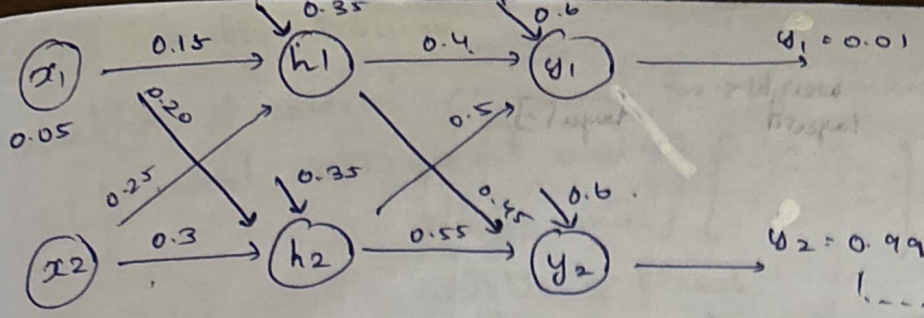
MSE / cross Entropy loss / Any other loss fn is Applicable

$\frac{\partial L}{\partial w_1}$, $\frac{\partial L}{\partial b_1}$

$\frac{\partial L}{\partial w_2}$



note:- In future we can replace the whole sigmoid eqn with one sigmoid gate Cause we know the derivative of sigmoid



Two output

$$\text{Loss} = \sum \frac{1}{2}(\text{target} - \text{output})^2$$

for all neurons in o/p layer.

$$h_1 = x_1 w_{11}^{[1]} + x_2 w_{12}^{[1]} + b_1^{[1]}$$

$$h_2 = x_1 w_{21}^{[1]} + x_2 w_{22}^{[1]} + b_2^{[1]}$$

$$y_1 = s_1 w_{11}^{[2]} + s_2 w_{12}^{[2]} + b_1^{[2]}$$

$$y_2 = s_1 w_{21}^{[2]} + s_2 w_{22}^{[2]} + b_2^{[2]}$$

s_1 = sigmoid(h_1)

s_2 = sigmoid(h_2)

$$w^{[1]} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} = \begin{bmatrix} 0.15 & 0.25 \\ 0.2 & 0.3 \end{bmatrix}$$

$$w^{[2]} = \begin{bmatrix} 0.4 & 0.5 \\ 0.45 & 0.55 \end{bmatrix}$$

$$b^{[1]} = \begin{bmatrix} 0.35 \\ 0.35 \end{bmatrix}$$

$$b^{[2]} = \begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}$$

$h_1 = 0.05 \times 0.15 + 0.10 \times 0.25 + 0.35$

$= 0.3825$

$h_2 = 0.05 \times 0.20 + 0.10 \times 0.3 + 0.35$

$= 0.39$

$s_1 = \frac{1}{1 + e^{-0.3825}} = 0.595$

$s_2 = \frac{1}{1 + e^{-0.39}} = 0.596.$

$y_1 = 0.595 \times 0.4 + 0.596 \times 0.5 + 0.6$

$\hat{y}_1 = 1.136$

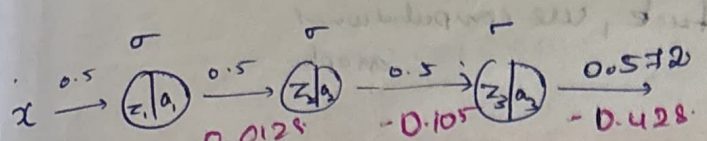
$y_2 = 0.595 \times 0.45 + 0.596 \times 0.55 + 0.6$

$\hat{y}_2 = 1.195$

$s\hat{y}_1 = 0.757$

$s\hat{y}_2 = 0.767$

Example :-



$z_1 = 0.5x = 0.5$

$a_1 = \sigma(z_1)$

$= \sigma(0.5x)$

$= \sigma(0.5)$

$= \frac{1}{1+e^{-0.5}} = 0.622$

$z_2 = 0.622 \times 0.5$ (a_1)

$= 0.311$

$a_2 = \sigma(z_2)$

$= \frac{1}{1+e^{-0.311}} = 0.577$

$z_3 = 0.577 \times 0.5 = 0.288$ (a_2)

$a_3 = \frac{1}{1+e^{-0.288}} = 0.572$

$= \frac{1}{1+e^{-z_3}}$

need to do with

$\frac{\partial L}{\partial a_3} \times \frac{\partial a_3}{\partial z_3}$

Exploding Gradients

↓

by ReLU

solved by clipping the gradient

[basically thresholding]

ground truth.

$y = 1$

$\frac{\partial a_3}{\partial a_3} = 1$

$\frac{\partial a_3}{\partial z_3} = \left(\frac{1}{1+e^{-z_3}}\right)\left(1 - \frac{1}{1+e^{-z_3}}\right)$

$= 2.75(1 - 2.77)$

$L = \frac{1}{2}(a_3 - y)^2$

① $\frac{\partial L}{\partial a_3} = a_3 - y = 0.572 - 1$

$= -0.428$

②

$\frac{\partial L}{\partial z_3} = \frac{\partial \frac{1}{2}\left(\frac{1}{1+e^{-z_3}} - y\right)^2}{\partial z_3}$

$= \left(\frac{1}{1+e^{-z_3}} - y\right)$ (1+2.75)

$= (2.75 - 1)(2.75)$

$(1 - 2.75)$

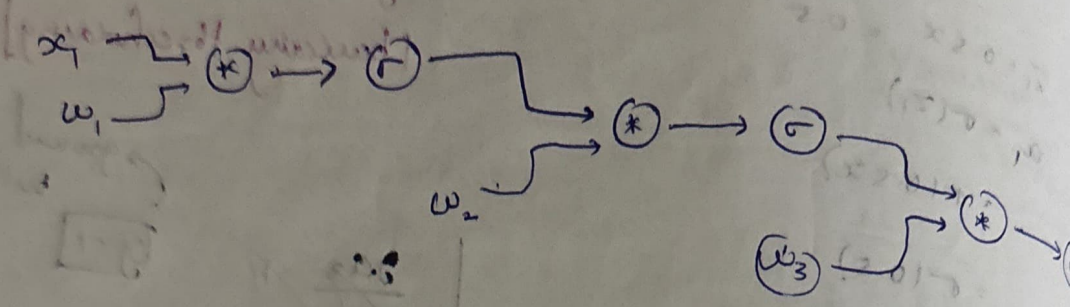
$= \frac{1}{(1+e^{-0.57})^{-1}}$

$= (0.571 - 1)(0.571)(1 - 0.57)$

$= -0.105$

(3) $\frac{\partial L}{\partial a_2} = \frac{\partial L}{\partial a_3} \frac{\partial a_3}{\partial z_3} \frac{\partial z_3}{\partial a_2} = -0.105 \times \frac{\partial 0.5 a_2}{\partial a_2} = -0.$

→ getting stuck, use computational graph



$$\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial a_3} \frac{\partial a_3}{\partial z_3} \frac{\partial z_3}{\partial a_2} \frac{\partial a_2}{\partial z_2}$$

$$= -0.0525 \times \frac{\partial \frac{1}{1+e^{-z_2}}}{\partial z_2}$$

$$= -0.0525 \times g(z_2)(1-g(z_2))$$

$$= -0.0525 \times 0.577 \times 0.423$$

$$= -0.012 = -0.013$$

$$\frac{\partial L}{\partial a_1} = \frac{\partial L}{\partial a_3} \frac{\partial a_3}{\partial z_3} \frac{\partial z_3}{\partial a_2} \frac{\partial a_2}{\partial z_2} \frac{\partial z_2}{\partial a_1}$$

$$= -0.013 \times 0.5 = -0.0064$$

$$\frac{\partial L}{\partial z_1} = -0.0064 \times \frac{\partial a_1}{\partial z_1}$$

$$= -0.0064 \times g(z_1)(1-g(z_1))$$

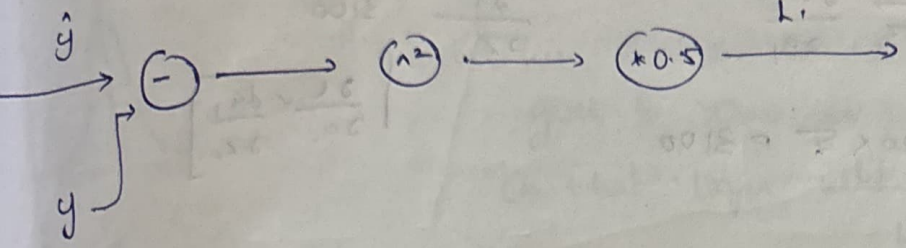
$$= -0.0064 \times 0.622 \times 0.378$$

$$= -0.001$$

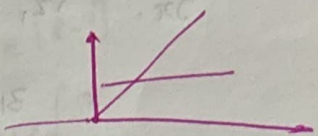
$$\frac{\partial L}{\partial x} = -0.005 \times \frac{\partial z_1}{\partial x}$$

$$= -0.0015 \times 0.5$$

$$= -0.00075 \longrightarrow \text{vanishing gradients.}$$



Exploding gradients



$$\frac{\partial L}{\partial z_3} = 125$$

$$\frac{\partial L}{\partial z_2} = 625$$

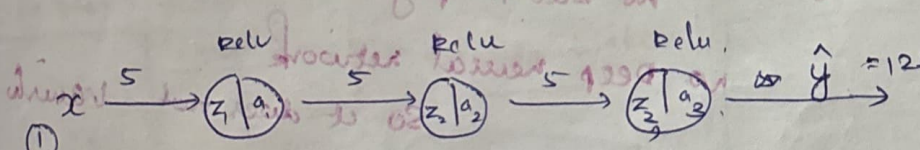
$$\frac{\partial L}{\partial z_1} = 3125$$

$$z_1 = wx = 5 \quad (w = 5,\ x = 1)$$

$$a_1 = \text{Relu}(z_1) = 5$$

$$z_2 = 5 a_1 = 5 \cdot 5 = 25$$

$$a_2 = \text{Relu}(z_2) = 25$$



$$z_1 = 5x = 5$$

$$a_1 = \text{Relu}(5) = 5$$

$$z_2 = 5 \times 5 = 25$$

$$a_2 = \text{Relu}(25) = 25$$

$$z_3 = 25 \times 5 = 125$$

$$a_3 = \text{Relu}(125) = 125$$

$$\boxed{a_3 = \hat{y} = 125}$$

$$L = \frac{1}{2}(\hat{y} - y)^2 = \frac{1}{2}(125 - 1)^2 = 7688$$

$$\frac{\partial L}{\partial a_3} = (a_3 - y) = 124$$

$$\frac{\partial L}{\partial z_3} = \frac{\partial L}{\partial a_3} \frac{\partial a_3}{\partial z_3} = 124 \frac{\partial \max(0, z_3)}{\partial z_3}$$

$$= 124 \times (124) = 124$$

$$\frac{\partial L}{\partial a_2} = \frac{\partial L}{\partial z_3} \times \frac{\partial z_3}{\partial a_2}$$

$$= 124 \times 5$$

$$= 620$$

$$\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial a_2} \frac{\partial a_2}{\partial z_2}$$

$$= 620 \times \frac{\partial \max(0, z_2)}{\partial z_2}$$

$$\frac{\partial L}{\partial z_2} = 620 \times 1$$

$$\frac{\partial L}{\partial a_1} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial a_1}$$

$$= 620 \times 5 = 3100$$

$$\frac{\partial L}{\partial z_1} = 3100$$

$$\left[\frac{\partial L}{\partial a_1} \times \frac{da_1}{\partial z_1} \right]$$

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z_1} \times \frac{\partial z_1}{\partial x}$$

$$= 3100 \times 5 = 15500$$

exploding gradient.

Even if exploding gradient we prefer ReLU because it easier to handle.

why not found in RL
no backpropagation
no deep neural network
so it does not vanish

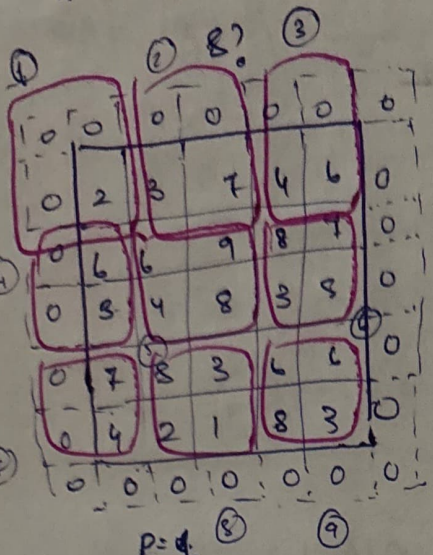
$$\|g\| = \sqrt{\sum_i g_i^2}$$

if $\|g\| > c$

$$g_{clipped} = g \times \frac{c}{\|g\|}$$

14/08/25 → Theory [Theory]

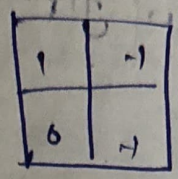
$n+2p-f+1$



p = 1

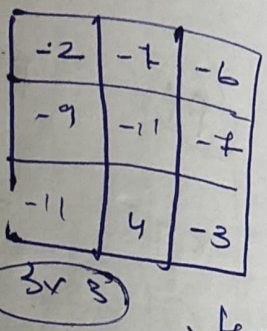
s = 2.

n = 5x5



f = 2x2.

⇒



3x3

floor

$n+2p-f+1//s$

$5+2-2+1 = 6//2$

but I need to get 3x3

3x3

↓

$\left[\frac{n-2p-f}{s}+1 \times \frac{n-2p-f}{s}+1\right]$

O/P.

(1) outside

(1) 1x0 + -1x0 + 0x0 + -1x2

(-2)

(2) (-7)

(3) (-6)

(4) -6 - 3

(-9)

(5) 6 - 9 - 8

-3 - 8

(-11)

(6) ~~8~~ - 7 - ~~8~~

(-7)

(7) -7 - 4

(-11)

(8) 8 - 3

-1

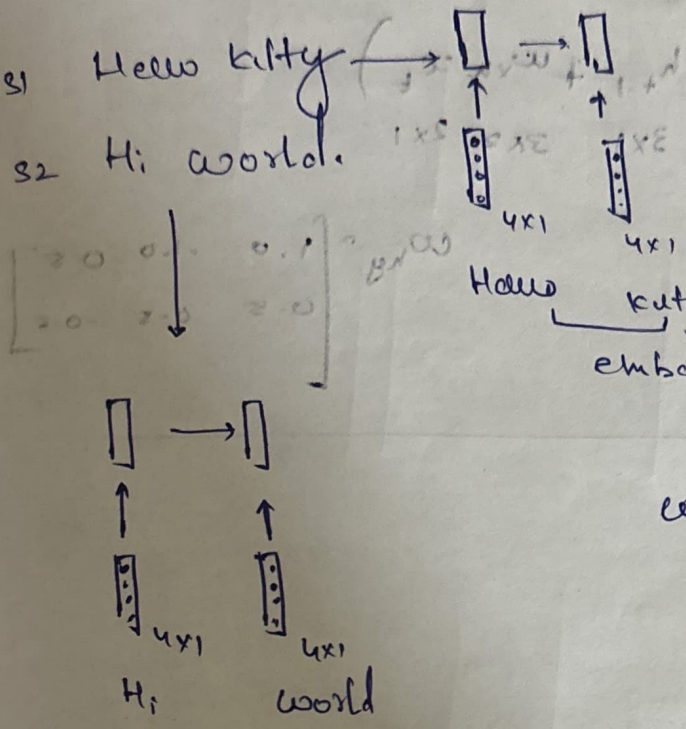
(4)

(9) 6 - 6 - 3

(-3)

Note :- W_{xh}, W_{hh} → Shared across the timesteps.

Note :- I/P change of length (two of coords changing in each sample) → handled by Recurrence by RNN block.



Why tanh :-
- Zero centered ops
- Bounded op stabilizes gradients.
- deriv @ all points
- gradient flow for temp dependencies

Why tanh as activation not Relu?

I/P to RNN always a vector, use a ResNet to get vector & then do RNN.

Character level Language Model

Predict the next character.

Vanilla RNN.

T = 2 (two time steps)

$X^{(1)} : x_1 \in \mathbb{R}^2 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$; $x_2 \in \mathbb{R}^2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Input dim = 2. $\Rightarrow x_t \in \mathbb{R}^2$

hidden state dim = 3 $\Rightarrow h_t \in \mathbb{R}^3$.

$h_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times 1}$

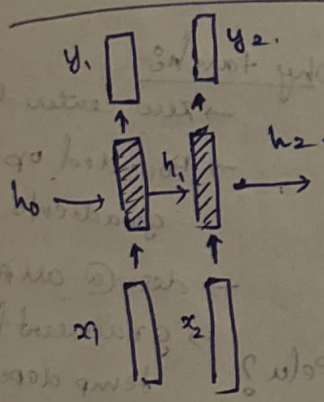
$y_t \in \mathbb{R}^2$.

$$W_{xh} = \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}_{3\times 2}, \quad W_{hh} = \begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix}_{3\times 3}$$

$$h_t = \tanh\left(\underset{3\times 3}{W_{hh}} \underset{3\times 1}{h_{t-1}} + \underset{3\times 2}{W_{xh}} \underset{2\times 1}{x_t} \right)$$

$$W_{hy} = \begin{bmatrix} 1.0 & -1.0 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix}$$

$$\underset{2\times 1}{y_t} = \underset{2\times 3}{W_{hy}} \underset{3\times 1}{h_t}$$



① forward pass

$$h_t = \tanh\left(\begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \right.$$

$$\begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix} \left(\begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$

$$h_t = \tanh\left([0 \quad 0 \quad 0] + [0.5 \quad -0.6 \quad 0.8 \quad 0.4 \quad 0.1 \quad -0.8]^T\right)$$

$$h_{t-1} = \tanh\left([-0.1 \quad 1.2 \quad 0.9]\right)$$

$$\tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$= \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

$$\tanh([-0.1 \quad 1.2 \quad 0.9])$$

① $= \begin{bmatrix} -0.221 \\ 2.221 \end{bmatrix} \begin{bmatrix} -0.099 \\ \end{bmatrix}$

② $= \begin{bmatrix} 0.571 \\ 1.449 \end{bmatrix} \begin{bmatrix} -0.380 & 0.83 \end{bmatrix}$

③ $\begin{bmatrix} -0.604 & 0.716 \end{bmatrix}$

$$h_t = \begin{bmatrix} -0.099 \\ -0.380 \\ 0.83 \\ -0.604 \\ 0.716 \end{bmatrix}_{3\times 1}$$

$$y_t = \begin{bmatrix} 1 & -1 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix} \begin{bmatrix} -0.099 \\ -0.380 \\ 0.83 \\ -0.604 \\ 0.716 \end{bmatrix}$$

$$= \begin{bmatrix} -0.099 - 0.380 - 0.302 \\ -0.045 - 0.19 + 0.302 \end{bmatrix}$$

$$= \begin{bmatrix} -0.781 \\ 0.067 \end{bmatrix}$$

$$y_1 = \begin{bmatrix} -0.571 \\ -0.48 \\ 0.08 \\ 0.0075 \end{bmatrix}_{2\times 1}$$

$$h_t = \tanh\left(w_{hh}\, h_{t-1} + w_{xh}\, y_{t-1}\right)$$

$$h_2 = \tanh\left(w_{hh}\, h_1 + w_{xh}\, y_1\right)$$

$$= \tanh\left(\begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix}\begin{bmatrix} -0.099 \\ 0.83 \\ 0.716 \end{bmatrix} + \begin{bmatrix} 0.5 & -0.3 \\ 0.0 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}\begin{bmatrix} -0.571 \\ 0.0078 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 0.3221 \\ 0.412 \\ 0.05525 \end{bmatrix} + \begin{bmatrix} -0.28775 \\ -0.4553 \\ -0.0541 \end{bmatrix}\right)$$

$$\tanh\left(\begin{bmatrix} 0.0643 \\ -0.043 \\ 0.00115 \end{bmatrix}\right)$$

$$h_2 = \begin{bmatrix} 0.0643 \\ -0.043 \\ 0.01138 \end{bmatrix} \qquad h_2 = \begin{bmatrix} -0.4434 \\ -0.1849 \\ 0.340 \end{bmatrix}$$

$$y_2 = \begin{bmatrix} -0.0888 \\ -0.4844 \end{bmatrix}$$

Corrected forward pass

$$h_t = \tanh\left(\begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix}\right)$$

$$h_1 = \begin{bmatrix} -0.099 \\ 0.83 \\ 0.716 \end{bmatrix}$$

$$\tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$y_1 = \begin{bmatrix} 1 & -1 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix}\begin{bmatrix} -0.099 \\ 0.83 \\ 0.716 \end{bmatrix}$$

$$y_1 = \begin{bmatrix} -0.571 \\ 0.0075 \end{bmatrix}_{2\times 1}$$

$$h_2 = \tanh\left(w_{hh} h_1 + w_{xh} x_2\right)$$

$$= \tanh\left(\begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix}\begin{bmatrix} -0.099 \\ 0.83 \\ 0.716 \end{bmatrix} + \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}\begin{bmatrix} -1 \\ 1 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 0.3321 \\ 0.42 \\ 0.0107 \end{bmatrix} + \begin{bmatrix} -0.8 \\ -0.6 \\ -0.3 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} -0.4779 \\ -0.1821 \\ -0.22 \\ 0.8107 \end{bmatrix}\right) = \begin{bmatrix} -0.4434 \\ -0.1847 \\ 0.801 \end{bmatrix}$$

24/09/2025

Note:-

LSTM minimises vanishing gradient cannot completely eliminate vanishing gradient

Why LSTM minimises vanishing gradient?

$$x_t = [0.5, -0.1]$$

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

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

$$w_{xg} = \begin{bmatrix} 0.5 & -0.3 \\ 0.4 & 0.1 \end{bmatrix}$$

$$w_{hi} = \begin{bmatrix} 0.1 & 0.2 \\ -0.2 & 0.05 \end{bmatrix}$$

$$w_{xf} = \begin{bmatrix} -0.4 & 0.2 \\ 0.3 & 0.3 \end{bmatrix}$$

$$w_{hf} = \begin{bmatrix} 0.05 & -0.1 \\ 0.2 & 0.1 \end{bmatrix}$$

$$w_{xo} = \begin{bmatrix} 0.3 & 0.2 \\ -0.2 & 0.2 \end{bmatrix}$$

$$w_{ho} = \begin{bmatrix} 0.15 & 0.05 \\ 0.1 & -0.2 \end{bmatrix}$$

$$w_{xg} = \begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix}$$

$$w_{hg} = \begin{bmatrix} 0.2 & 0.1 \\ -0.1 & 0.05 \end{bmatrix}$$

$$i_t = \sigma(w_{hi} h_{t-1} + w_{xi} x_t)$$

$$= \sigma\left(\begin{bmatrix} 0.1 & 0.2 \\ -0.2 & 0.05 \end{bmatrix}_{2\times2} \begin{bmatrix} 0 \\ 0.1 \end{bmatrix}^T + \begin{bmatrix} 0.5 & -0.3 \\ 0.4 & 0.1 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}_{1\times2}\right)$$

$$= \sigma\left(\begin{bmatrix} 0+0.02 \\ 0+0.005 \end{bmatrix} + \begin{bmatrix} 0.25+0.03 \\ 0.2-0.01 \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} 0.02 \\ 0.005 \end{bmatrix} + \begin{bmatrix} 0.28 \\ 0.19 \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} 0.30 \\ 0.195 \end{bmatrix}\right) = \begin{bmatrix} 0.57 \\ 0.55 \end{bmatrix}$$

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

$$= \tanh\left(\begin{bmatrix} 0.2 & 0.1 \\ -0.1 & 0.05 \end{bmatrix}\begin{bmatrix} 0 \\ 0.1 \end{bmatrix} + \begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix}\begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 0.01 \\ 0.005 \end{bmatrix} + \begin{bmatrix} -0.25-0.04 \\ 0.1+0.03 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 0.01 \\ 0.005 \end{bmatrix} + \begin{bmatrix} -0.29 \\ 0.13 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} -0.28 \\ 0.135 \end{bmatrix}\right) = \begin{bmatrix} -0.27 \\ -0.14 \\ 0.001 \\ 0.13 \end{bmatrix}$$

$$f(t) = \sigma(W_{hf} h_{t-1} + W_{xf} x_t)$$

$$= \sigma\left(\begin{bmatrix} 0.05 & -0.1 \\ 0.2 & 0.1 \end{bmatrix} \begin{bmatrix} 0 \\ 0.1 \end{bmatrix} + \begin{bmatrix} -0.4 & 0.2 \\ 0.3 & 0.3 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix} \right)$$

$$= \sigma\left(\begin{bmatrix} 0 & -0.01 \\ 0.01 & \end{bmatrix} + \begin{bmatrix} -0.2 & 0.02 \\ 0.15 & -0.03 \end{bmatrix} \right)$$

$$= \sigma\left(\begin{bmatrix} -0.01 \\ 0.01 \end{bmatrix} + \begin{bmatrix} -0.22 \\ 0.12 \end{bmatrix} \right)$$

$$= \sigma\left(\begin{bmatrix} -0.23 \\ 0.13 \end{bmatrix} \right)$$

$$f(t) = \begin{bmatrix} 0.443 \\ 0.53 \end{bmatrix}$$

element wise multiplication

$$C(t) = f_t \odot C_{t-1} + i_t \odot g_t$$

$$= \begin{bmatrix} 0.443 \\ 0.53 \end{bmatrix} \odot \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix} + \begin{bmatrix} 0.57 \\ 0.58 \end{bmatrix} \begin{bmatrix} -0.27 \\ 0.13 \end{bmatrix}$$

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

$$= \begin{bmatrix} 0.0886 \\ -0.106 \end{bmatrix} + \begin{bmatrix} -0.154 \\ 0.075 \end{bmatrix}$$

$$= \begin{bmatrix} -0.0654 \\ -0.031 \end{bmatrix}$$

here $o_t = \sigma(w_{ho} h_{t-1} + w_{xo} x_t)$

$$o_t = \sigma\left(\begin{bmatrix} 0.15 & 0.05 \\ 0.1 & -0.2 \end{bmatrix} \begin{bmatrix} 0 \\ 0.1 \end{bmatrix} + \begin{bmatrix} 0.3 & 0.2 \\ -0.2 & 0.2 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix} \right)$$

$$= \sigma\left(\begin{bmatrix} 0.05 \\ -0.02 \end{bmatrix} + \begin{bmatrix} 0.15 & -0.02 \\ 0.1 & -0.02 \end{bmatrix} \right)$$

$$= \sigma\left(\begin{bmatrix} 0.05 \\ -0.02 \end{bmatrix} + \begin{bmatrix} 0.13 \\ -0.12 \end{bmatrix} \right)$$

$$= \sigma\left(\begin{bmatrix} 0.18 \\ -0.14 \end{bmatrix} \right) = \begin{bmatrix} 0.55 \\ 0.46 \end{bmatrix}$$

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

$$= \begin{bmatrix} 0.55 \\ 0.46 \end{bmatrix} \cdot \tanh \begin{bmatrix} -0.0654 \\ -0.031 \end{bmatrix} = \begin{bmatrix} 0.55 \\ 0.46 \end{bmatrix} \begin{bmatrix} -0.065 \\ -0.03 \end{bmatrix}$$

$$h_t = \begin{bmatrix} -0.03575 \\ -0.014 \end{bmatrix}$$

Sum

$h_1 = [1, 0, 0]$

$h_2 = [0, 1, 1]$

$h_3 = [1, 1, 0]$

$t = 3$

$s_{t-1} = [1, 0, 1]$

Score fn = dot product

$c_t = ?$

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

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

$$\alpha_{t,1} = \frac{\exp(\text{score}(s_{t-1}, h_j))}{\sum_{j=1}^{T} \exp(\text{score}(s_{t-1}, h_j))}$$

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

$S_{t-1} \cdot h_2 = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$

$= [1 + 0 + 0]$

$S_{t-1} \; h_2 = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$

$S_{t-1} \; h_1 = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$

$= 2$

$\alpha_{t,1} = \dfrac{\exp(2)}{\exp(1) + \exp(1) + \exp(2)} = 0.576$

$\alpha_{t,2} = \dfrac{\exp(1)}{\exp(1) + \exp(1) + \exp(2)} = 0.212$

$\alpha_{t,3} = 0.212$

$C_t \rightarrow$ vector

$= \begin{bmatrix} 0.576 \\ 0 \\ 0.576 \end{bmatrix} + \begin{bmatrix} 0 \\ 0.212 \\ 0.212 \end{bmatrix} + \begin{bmatrix} 0.212 \\ 0.212 \\ 0 \end{bmatrix}$

$= \begin{bmatrix} 0.788 \\ 0.424 \\ 0.788 \end{bmatrix}$ → more closer to h_1

that's how the attention mechanism works.

How does chatGPT work?

How does . .

→ This is exactly similar to Attention

Transformers are used in this. I can be used as a database retrieval problem

S_{t-1} → Query

key	value
[1.0 0.5 0.2]	"Action movie Die hard"
[0.3 1.0 0.1]	"Romance Titanic"
[0.9 0.4 0.8]	"Action movies John Wick"

Q ⇒ looking for a action movie

↓ word2vec

[1.0, 0.3, 0.5]

$$\begin{bmatrix} 1.0 & 0.5 & 0.2 \end{bmatrix} \begin{bmatrix} 1 \\ 0.3 \\ 0.5 \end{bmatrix} = [1 + 0.15 + 0.10] = [1.25]$$

$$\begin{bmatrix} 0.3 & 1 & 0.1 \end{bmatrix} \begin{bmatrix} 1 \\ 0.3 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.3 + 0.3 + 0.05 \end{bmatrix}$$

$$\therefore \quad \begin{bmatrix} 0.65 \end{bmatrix}$$

$$\begin{bmatrix} 0.9 & 0.4 & 0.8 \end{bmatrix} \begin{bmatrix} 1 \\ 0.3 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.9 + 0.12 + 0.4 \end{bmatrix}$$

$$= \begin{bmatrix} 1.42 \end{bmatrix}$$

0.52
0.9
1.42

↓

Highest similarity

Transformers [encoder - Decoder network]

RNN disadv°-

→ long Range dependencies
→ vanishing gradient

] handle temporal Data

Removed recurrence → makes it parallelisable

→ Attention mechanism

→ Multi head Attention → self attention

→ Temporal Dependency lost?

Does not have recurrence

in order to fix that they introduced positional encoding

Diagram

Example 8:

I/P # Playing outside O/P : z_1 z_2 ?

embedding [ ] [ ] (word2vec)

$\times W^Q$

$\times W^K$

$\times W^V$

Playing : x_1 = [0.212 0.04 0.63 0.36]

q_1 [0.31 0.84 0.963 0.57]

k_1 [0.36 0.83 0.1 0.38]

v_1

outside : x_2 = [0.1 0.14 0.86 0.7]

q_2 = [0.45 0.94 0.73 0.58]

k_2 = [0.31 0.36 0.19 0.72]

v_2

Playing.

$\Rightarrow q_1^T \cdot k_1 = p_{11}$

$= [0.212 \quad 0.04 \quad 0.63 \quad 0.36] \begin{bmatrix} 0.31 \\ 0.84 \\ 0.963 \\ 0.57 \end{bmatrix}$

$= 0.91121 / \sqrt{4} = 0.4556$

$\Rightarrow q_1^T \cdot k_2 = p_{12}$

$= [0.212 \quad 0.04 \quad 0.63 \quad 0.36] \begin{bmatrix} 0.45 \\ 0.94 \\ 0.73 \\ 0.58 \end{bmatrix}$

$= 0.8017 / \sqrt{4} = 0.40085$

Softmax : $\frac{e^{-P_{11}}}{e^{-P_{11}} + e^{-P_{12}}}$ $\frac{e^{-P_{12}}}{e^{-P_{11}} + e^{-P_{12}}}$

$= \frac{0.634}{1.3038}$ $\frac{0.669 \text{ f}}{1.3038}$

0.486 0.513

Softmax ✓

$0.486 \begin{bmatrix} 0.36 \\ 0.13 \\ 0.1 \\ 0.38 \end{bmatrix} + 0.513 \begin{bmatrix} 0.31 \\ 0.36 \\ 0.14 \\ 0.712 \end{bmatrix}$

I_1 :

$\begin{bmatrix} 0.334 \\ 0.588 \\ 0.146 \\ 0.554 \end{bmatrix}$ $-$

Outside

$\Rightarrow a_2^T k_1 = P_{21}$

$= [0.1 \quad 0.14 \quad 0.86 \quad 0.77] \begin{bmatrix} 0.31 \\ 0.84 \\ 0.963 \\ 0.57 \end{bmatrix}$

$= 1.41568 / 2 = 0.70784$

$\Rightarrow q_2^T k_2$

$$\begin{bmatrix} 0.1 & 0.14 & 0.86 & 0.77 \end{bmatrix} \begin{bmatrix} 0.31 \\ 0.36 \\ 0.19 \\ 0.72 \end{bmatrix}$$

$= 0.7912/2 = 0.3996$

$$\text{Softmax} \begin{bmatrix} \frac{e^{-0.70784}}{e^{-0.70784} + e^{-0.3996}} & \frac{e^{-0.3996}}{e^{-0.70784} + e^{-0.3996}} \end{bmatrix}$$

$$= \frac{0.493}{1.1633} \qquad \frac{0.6706}{1.1633}$$

$$= 0.4238 \qquad 0.5765$$

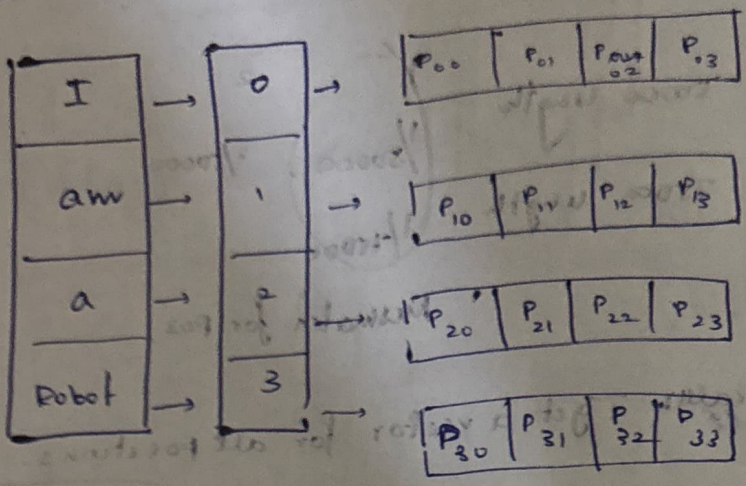
$$= 0.4238 \begin{bmatrix} 0.36 \\ 0.83 \\ 0.1 \\ 0.38 \end{bmatrix} \quad 0.5765 \begin{bmatrix} 0.31 \\ 0.36 \\ 0.19 \\ 0.72 \end{bmatrix}$$

$$Z_2 = \begin{bmatrix} 0.3812 \\ 0.5593 \\ 0.151915 \\ 0.5612 \end{bmatrix}$$

$0 < i < d/2$

I am a Robot.

$P_E : d = 4 \Rightarrow \mathbb{R}^4$



$i = 0$:-

P_{00}
P_{01}

$$P_{pos,2i} = \sin\left(\frac{pos}{10000^{2i}}\right)$$

$$P_{pos,2i+1} = \cos\left(\frac{pos}{10000^{2i}}\right)$$

$$P_{00} = \sin\left(\frac{0}{10000^{0}}\right)$$

$$= \sin 0$$

$$= 0$$

$$P_{01} = \cos\left(\frac{0}{10000^{0/4}}\right)$$

$$P_{01} = 1$$

$i = 1$:-

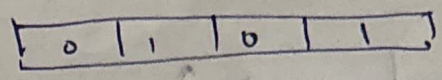
P_{02}
P_{03}

$$P_{02} = \sin\left(\frac{0}{10000^{2/4}}\right)$$

$$= 0$$

$$P_{03} = \cos\left(\frac{0}{10000^{2/4}}\right)$$

$$= 1$$



i = 0.

P_{10}
P_{11}]

$$P_{10} = \sin\left(\frac{1}{10000^{2\times 0/4}}\right)$$

$$= \sin\left(\frac{1}{10000^{0}}\right)$$

$$= \sin 1 = 0.0175$$

$$P_{11} = \cos\left(\frac{1}{10000^{2\times 0/4}}\right)$$

$$= \cos 1 = 0.9998.$$

i = 1

P_{12}
P_{13}]

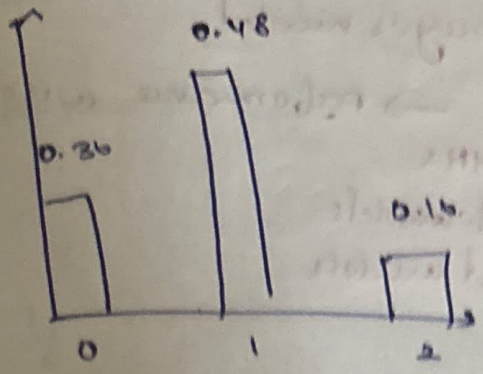
$$P_{12} = \sin\left(\frac{1}{10000^{2\times 1/4}}\right)$$

$$= \sin\left(\frac{1}{100}\right) = 0.00145$$

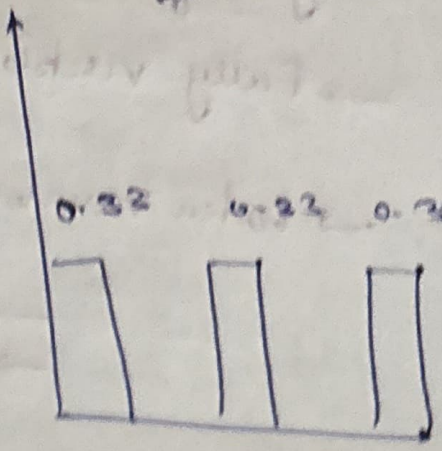
$$P_{13} = \cos\left(\frac{1}{100}\right) = 0.9999$$

KL example

P. Binomial distbn

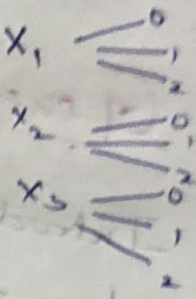


Q uniform distbn



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

for bunch of samples.



[0, 1, 2]
bunch of R.V

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

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

$$\approx 0.0852996$$

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

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

$$= -0.0256 + (-0.1215) + 0.24465 = 0.0975$$