$= W^{[3]} W^{[2]} g(z^{[1]}) + b^{[2]} + b^{[3]}$.

$= W^{[3]} W^{[2]} z^{[1]} + b^{[2]} + b^{[3]}$.

$= (W^{[3]} W^{[2]} W^{[1]}) a^{[0]} + (b^{[1]} + b^{[2]} + b^{[3]})$

$a^{[0]} \Rightarrow x$

$\tilde{W} \Rightarrow W$

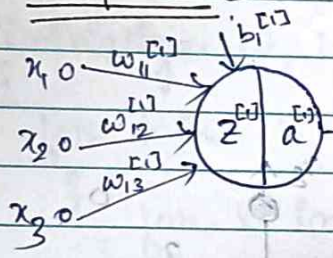
$\tilde{W} x + b$

Representation power of NN.

→ NN with atleast one hidden layer are universal approximators.

→ For any continuous func^n, $f(x)$, and some $\epsilon > 0$, there exists a NN-$g(x)$ with atleast 1 hidden layer (with some non-linearity) such that $\forall x \; |f(x) - g(x)| \leq \epsilon$

Perceptron

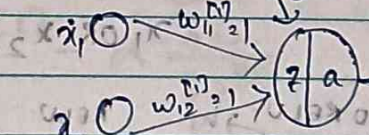


$z^{[1]} = w_{11}^{[1]} x_1 + w_{12}^{[1]} x_2 + w_{13}^{[1]} x_3 + b^{[1]}$

$a^{[1]} = g(z^{[1]})$. $g = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z \leq 0 \end{cases}$

Logical AND

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



For $x_1 = 0$ & $x_2 = 0$.

$z = 1 \times 0 + 1 \times 0 + (-1) = -1$.

$a = g(z) = 0$. Output $= 0 = y$.

For $x_1 = 0$ & $x_2 = 1$.

$z = 1 \times 0 + 1 \times 1 + (-1) = 0$.

$a = g(z) = 0$, $\hat{y} = 0 = y$.

For $x_1 = 1$ & $x_2 = 0$.

$z = 1 \times 1 + 1 \times 0 + (-1)$

$a = g(z) = 0$, $\hat{y} = 0 = y$

Logical OR

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

apsara

Date: / /

For $x_1 = x_2 = 1$

$z = 1 \times 1 + 1 \times 1 + (-1)$

$= 1.$

$a = g(z) = \hat{y} = 1 = y.$

when $b = 0$, logical OR works for perceptron.

XOR table.

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

For $x_1 = 1, x_2 = 1$.

$z = 1 \times 1 + 1 \times 1 + 0$

$= 2.$

$a = 1 = \hat{y} \neq y.$

For $x_1 = x_2 = 0$.

$z = 1 \times 0 + 1 \times 0 + 0.$

$= 0.$

$a = 0 = \hat{y}.$

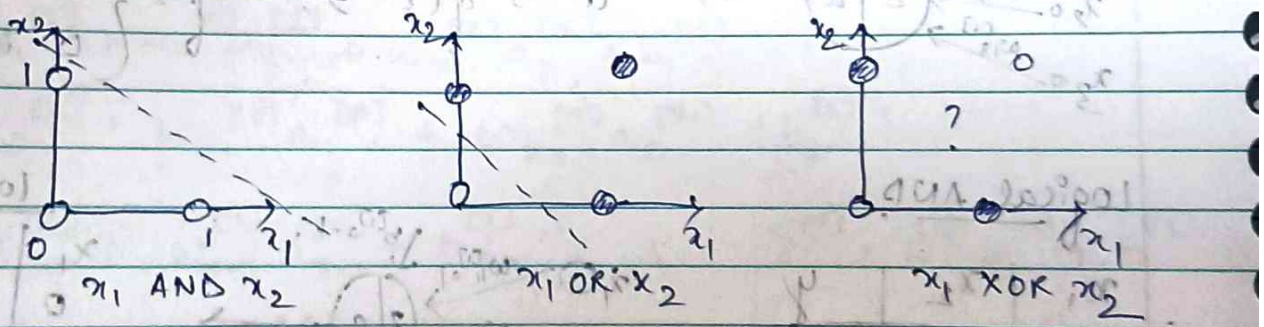
For $x_1 = 1, x_2 = 0$

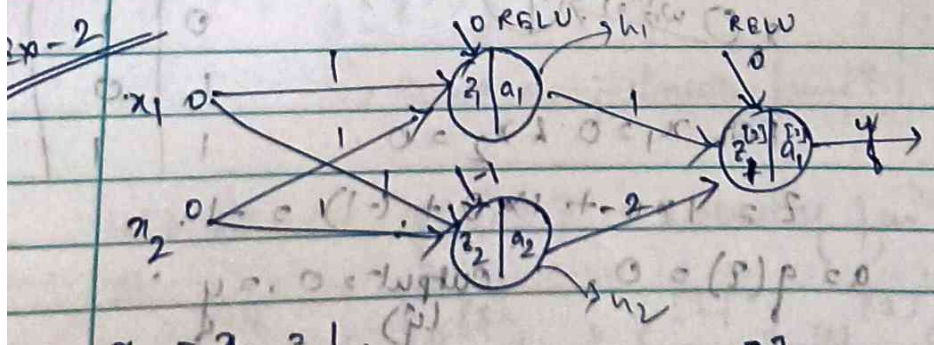
$z = 1 \times 1 + 1 \times 0 + 0.$

$= 1$

$a = 1 = \hat{y}.$

Decision boundaries.





$x_1 = x_2 = 1.$

$z_1^{[1]} = 1 \times 1 + 1 \times 1 + 0 = 2.$ $a_1^{[1]} = \max(0, 2) = 2.$

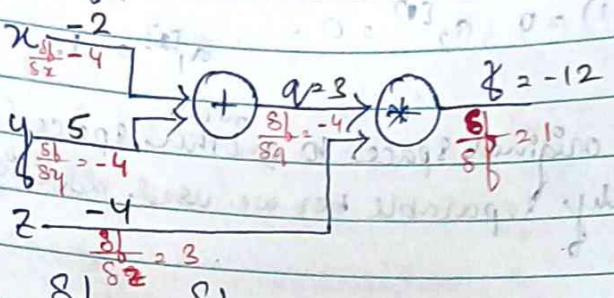
$z_2^{[1]} = 1 \times 1 + 1 \times 1 + (-1) = 1.$ $a_2^{[1]} = \max(0, 1) = 1.$

$z_1^{[2]} = 2 \times 1 + 1 \times (-2) + 0 = 0.$ $a_1^{[2]} = \max(0, 0) = 0.$

Page No.:

Date 29 07 25

$f = (x+y)z$.



let $q = x+y$

so, $f = qz$.

$f = (-2+5)\cdot(-4)$

$= -12$.

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

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

$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} = ?$

$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial x}$ → local gradient.

↳ global gradient.

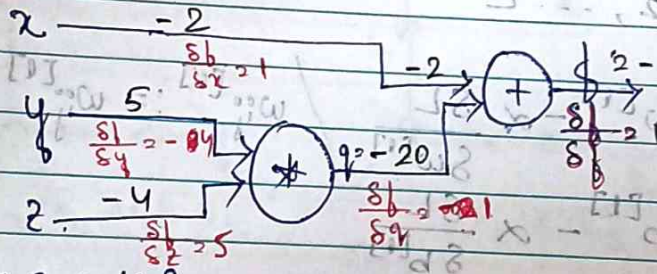
$= -4 \times 1 = -4$.

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

$= -4 \times 1 = -4$

Q4 → $f(x,y,z) = x + yz$.

$x = -2, y = 5, z = -4$.



let $q = yz$.

so, $f = x + q$. → $\frac{\partial f}{\partial x} = 1, \frac{\partial f}{\partial q} = 1$

$f = -2 - 20 = -22$.

$\frac{\partial q}{\partial y} = z, \frac{\partial q}{\partial z} = y$.

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

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

apsara

Date 30.07.25

Steps:-

1. Identify additional functions
2. Draw a computational graph.
3. Compute forward pass.
4. Perform backward pass starting from end of the circuit.

eg.2 $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$

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

① $b_3 = w_0x_0 + w_1x_1 + w_2 = b_1 + b_2 + w_2$

$f(w,x) = \frac{1}{1+e^{-z}}$ $b_4 = -b_3, b_5 = e^{b_4}$

$b_6 = 1 + b_5$

$b_7 = 1/b_6$

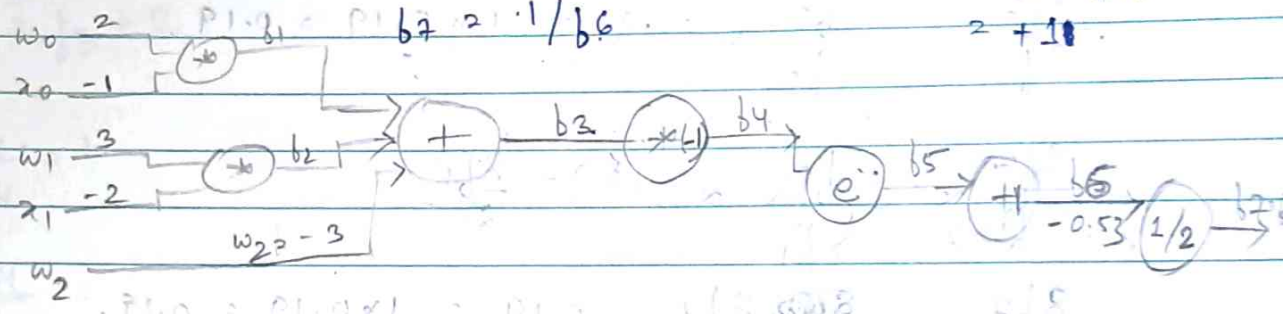
$b_1 = w_0x_0 = 2 \times -1 = -2$

$b_2 = w_1x_1 = -3 \times (-2) = +6$

$b_3 = b_1 + b_2 + w_2$

$= -2 + (6) + (-3)$

$= +1$

② 

$b_4 = (-1)$, $b_5 = e^{-1} = 0.36$

$b_6 = 1 + 0.36 = 1.36$

$b_7 = \frac{1}{1.36} = 0.73$.

$\frac{\delta b_7}{\delta b_7} = 1$, $\frac{\delta b_1}{\delta w_0} = x_0$, $\frac{\delta b_1}{\delta x_0} = w_0$.

Local gradients:-

$\frac{\delta b_1}{\delta w_0} = x_0 = -1$	$\frac{\delta b_2}{\delta x_1} = w_1 = -3$
$\frac{\delta b_1}{\delta x_0} = w_0 = 2$	$\frac{\delta b_3}{\delta b_1} = 1$
$\frac{\delta b_2}{\delta w_1} = x_1 = -2$	$\frac{\delta b_3}{\delta b_2} = 1$
	$\frac{\delta b_3}{\delta w_2} = 1$

Global gradients:-

$\frac{\delta b_6}{\delta b_5} = 1$

$\frac{\delta b_5}{\delta b_4} = e^{b_4} = e^{0.36}$

$\frac{\delta b_4}{\delta b_3} = -1$

$\frac{\delta b_7}{\delta b_6} = -\frac{1}{b_6^2}$

$= -\frac{1}{1.36^2} = -0.53$

Page No.:

apsara

Date:

Global gradients :-

$$\frac{\delta f_7}{\delta f_6} = \frac{-1}{b_6^2} \times 1 = \frac{-1}{1.36^2} \times 1 = -0.54$$

$$\frac{\delta f_7}{\delta f_5} = \frac{\delta f_6}{\delta f_5} \times (-0.54) = 1 \times (-0.54) = -0.54$$

$$\frac{\delta f_7}{\delta f_4} = \frac{\delta f_5}{\delta f_4} \times (-0.54) = e^{-1} \times (-0.54) = -0.19$$

$$\frac{\delta f_7}{\delta f_3} = \frac{\delta f_4}{\delta f_3} \times (-0.19) = -1 \times (-0.19) = 0.19$$

$$\frac{\delta f_7}{\delta f_2} = \frac{\delta f_3}{\delta f_2} \times 0.19 = 1 \times 0.19 = 0.19$$

$$\frac{\delta f_7}{\delta w_2} = \frac{\delta f_3}{\delta w_2} \times 0.19 = 1 \times 0.19 = 0.19$$

$$\frac{\delta f_7}{\delta f_2} = \frac{\delta f_3}{\delta f_2} \times 0.19 = 1 \times 0.19 = 0.19$$

$$\frac{\delta f_7}{\delta f_1} = \frac{\delta f_3}{\delta f_1} \times 0.19 = 1 \times 0.19 = 0.19$$

$$\frac{\delta f_7}{\delta w_0} = \frac{\delta f_1}{\delta w_0} \times 0.19 = -1 \times 0.19 = -0.19$$

$$\frac{\delta f_7}{\delta x_0} = \frac{\delta f_1}{\delta x_0} \times 0.19 = 2 \times 0.19 = 0.38$$

$$\frac{\delta f_7}{\delta x_1} = \frac{\delta f_2}{\delta x_1} \times 0.19 = -3 \times 0.19 = -0.57$$

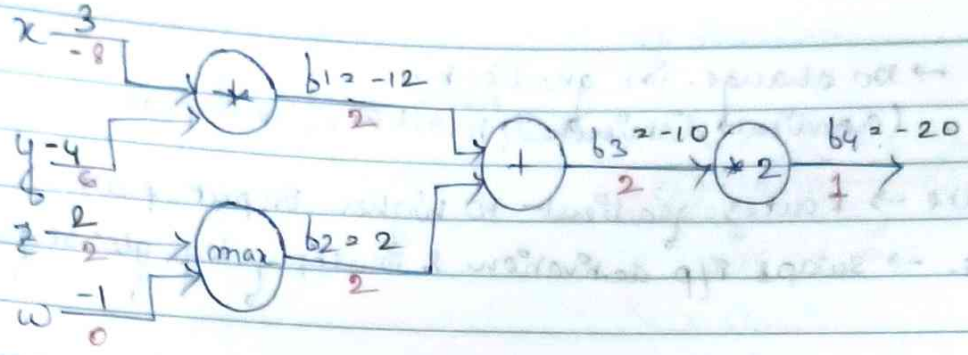
$$\frac{\delta f_7}{\delta w_1} = \frac{\delta f_2}{\delta w_1} \times 0.19 = -2 \times 0.19 = -0.38$$

Page No.:

apsara

Date:

q.3



$b_1 = x * y$

$b_2 = \max(z, w)$

$b_3 = b_1 + b_2$

$b_4 = b_3 * 2$

$\frac{\delta b_1}{\delta x} = y = -4$

$\frac{\delta b_1}{\delta y} = x = 3$

$\frac{\delta b_2}{\delta w} =$

$\frac{\delta b_2}{\delta z} =$

$\frac{\delta b_3}{\delta b_1} = 1 \quad \delta$

$\frac{\delta b_3}{\delta b_2} = 1$

$\frac{\delta b_4}{\delta b_3} = 2.$

$\frac{\delta b_4}{\delta b_4} = 1$

$\frac{\delta b_4}{\delta b_3} = 2 \times 1 = 2.$

$\frac{\delta b_4}{\delta b_1} = \frac{\delta b_3}{\delta b_1} \times 2 = 1 \times 2 = 2$

$\frac{\delta b_4}{\delta x} = \frac{\delta b_1}{\delta x} \times 2 = -4 \times 2 = -8$

$\frac{\delta b_4}{\delta y} = \frac{\delta b_1}{\delta y} \times 2 = 3 \times 2 = 6.$

$\frac{\delta b_2}{\delta w} = 1$ if $w > z$.
0 if $w < z$.
undefined if $w = z$.

→ 0.

$\frac{\delta b_4}{\delta z} = \frac{\delta b_2}{\delta z} \times 2 = 1 \times 2 = 2$

$\frac{\delta b_4}{\delta w} = \frac{\delta b_2}{\delta w} \times 2 = 0 \times 2 = 0$

$\frac{\delta b_4}{\delta b_2} = \frac{\delta b_3}{\delta b_2} \times 2$

$= 1 \times 2 = 2.$

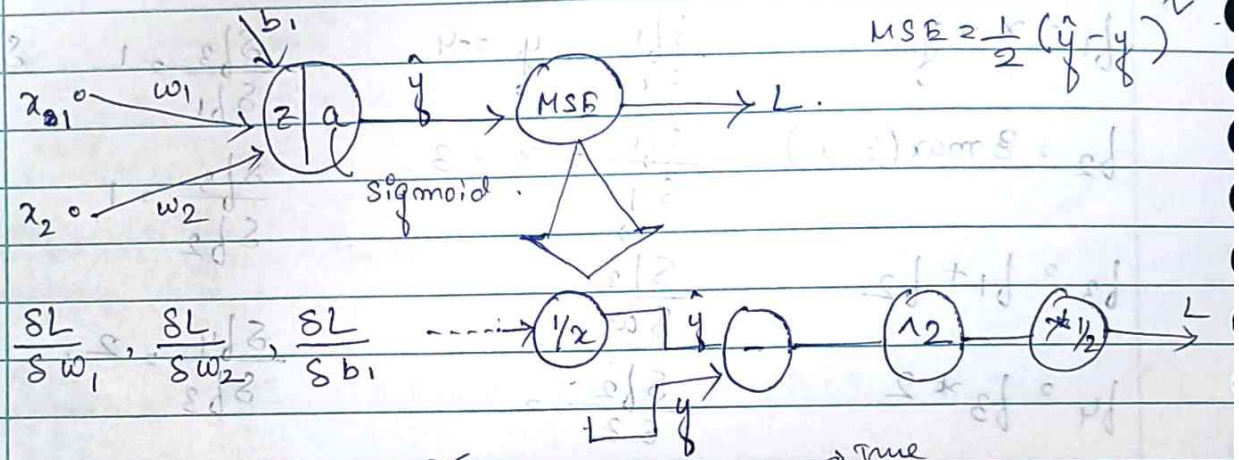
$\frac{\delta b_2}{\delta z} = 1$ if $z > w$.
0 if $z < w$.
undefined $z = w$.
arbitary.

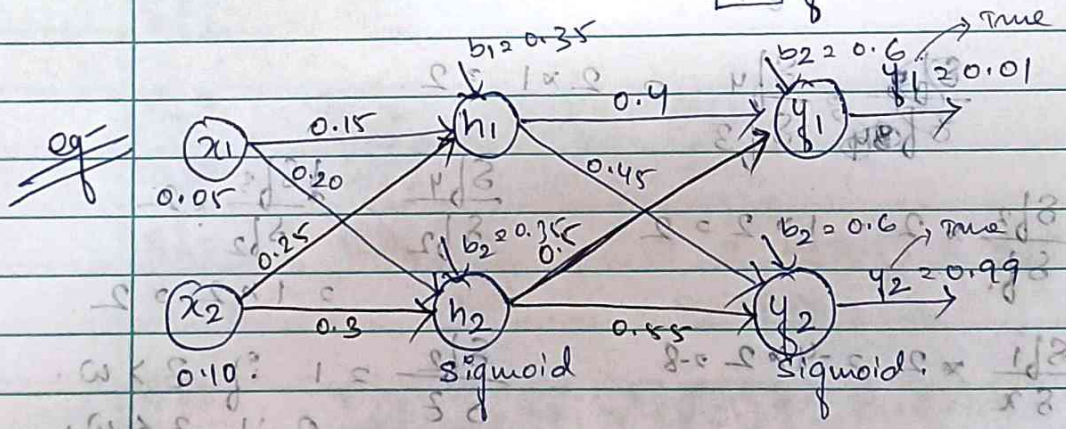
→ 1.

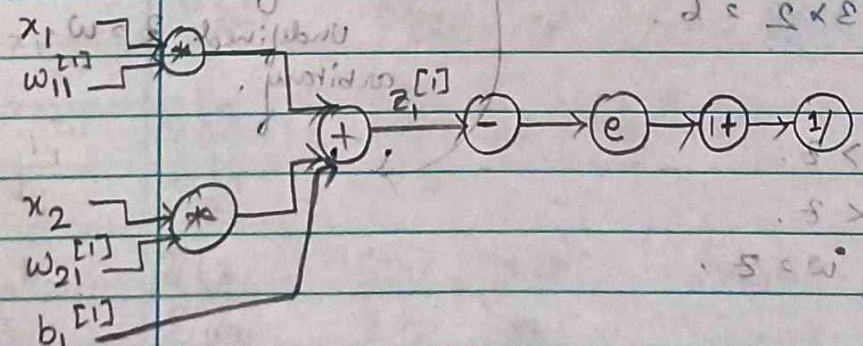
(+) gate → no change in gradient. (gradient distributed equally to all of its inputs.

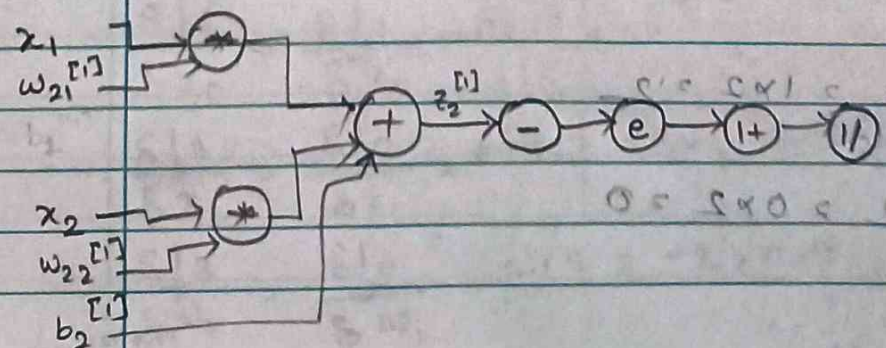
(max) gate → Routes gradient to higher input path.

(*) gate → swaps i/p activation & multiply by global gradient.







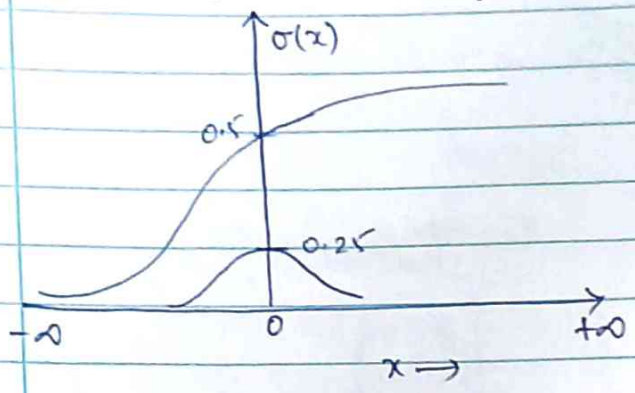


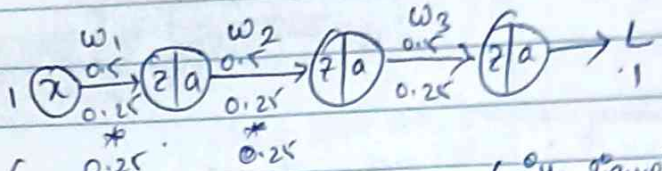
Page No.:

apsara

Date 04.09.25

Vanishing & Exploding Gradients Problem.

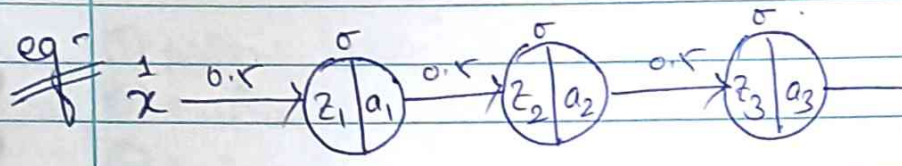




↳ vanishing gradient (with sigmoid activation fun)
(so no training / learning)
(Also for TanH activation fun)

To fix vanishing gradient problem → ReLU is used.
↓ but this leads to
Exploding gradient Problem
(clipping the gradient reduces this problem).

eg:



$z_3 = a_2 w_3$

$z_1 = 1 \times 0.5 = 0.5$
$a_1 = \sigma(0.5) = 0.622.$

$z_2 = 0.622 \times 0.5$
$= 0.311$
$a_2 = \sigma(0.311) = 0.577.$

$z_3 = 0.577 \times 0.5$
$= 0.2885$
$a_3 = \sigma(0.2885)$
$= 0.571.$

$\text{Loss}(L) = \frac{1}{2}(0.571 - 1)^2 = 0.092.$

$\frac{\delta L}{\delta w_3} = \frac{\delta L}{\delta a_3} \times \frac{\delta a_3}{\delta z_3} \times \frac{\delta z_3}{\delta w_3} = -0.428 \times 0.205 \times 0.577 =$

$\frac{\delta L}{\delta a_3} = (a_3 - y) = 0.571 - 1 = -0.428.$

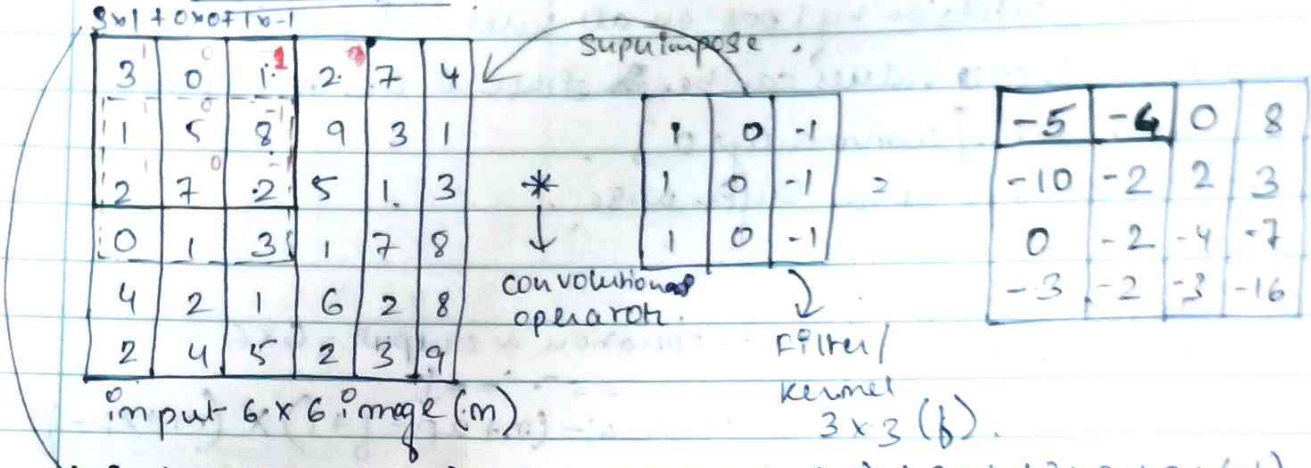
$\frac{\delta a_3}{\delta z_3} = 0.288(1 - 0.288) = 0.205$

$\frac{\delta z_3}{\delta w_3} = a_2 = 0.577.$

apsara

Date 13 08 25

Convolution operation :-



$3\times1 + 0\times0 + 1\times(-1) + 1\times1 + 5\times0 + 8\times(-1) + 2\times1 + 7\times0 + 2\times(-1)$

$= -5$

Slide filter by 1:

$1\times0 + 0\times1 + 2\times-1 + 5\times1 + 8\times(0) + 9\times(-1) + 7\times1 + 2\times0 + 5\times(-1)$

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

Vertical filter.

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

Horizontal filter.

Both for edge detection.

Edge detectors — Sobel edge dection.
Canny " "
Laplacian " "
Prewitt " "
Scharr " "

Output image size = $(n-f+1) \times (n-f+1)$

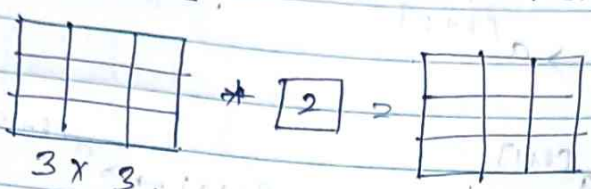
Padding

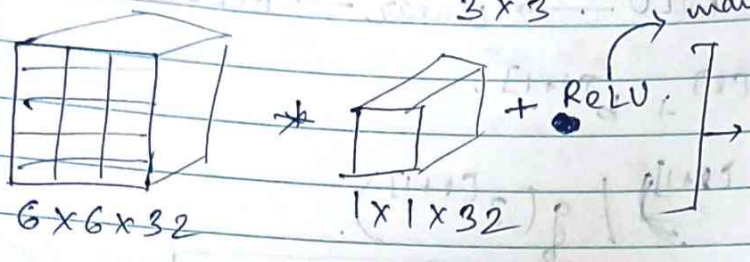
For every layer → input image will shrunk. (Just like 6×6 → 4)
Filter does not see all the pixels as same no. of times.
Problem.
Solve 2 problem —
① Shrinking problem. ② Equal attention to all pixels.

Page No.:

Date: 26.08.25

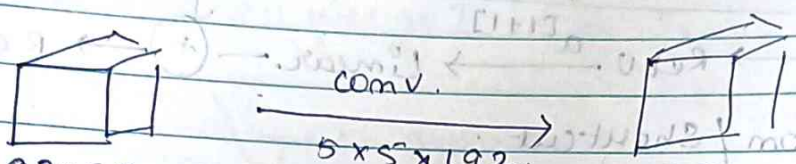
1 × 1 Conv / Network in a Network.





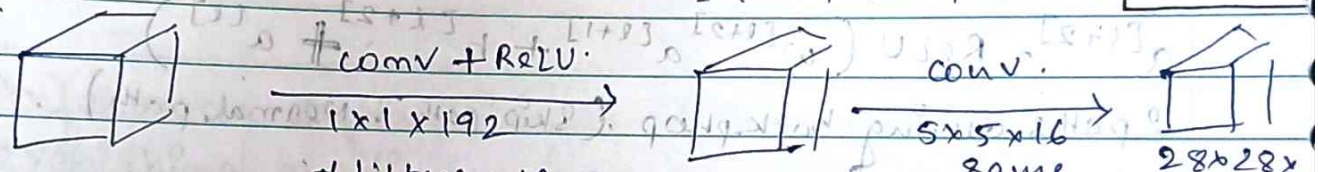
$n - f + 1$
$= 28 - 5 + 1$
$= 24$
$n - f$
3×3 * 2×2
16.
$3 \times 3 * 2 \times 2$
$9 * 4$
36

eg.1



Total operations (multiplications) = $5 \times 5 \times 192 \times 28 \times 28 \times 32$
$= 120,422,400 \approx$ **120 million**

eg.2



$28 \times 28 \times 16 \rightarrow 2408448 \approx 2.4$ million

$28 \times 28 \times 32 \rightarrow 10,035,200 \approx 10$ million

$2408448 + 10,035,200 = 12443648 \approx$ **12 million**

(*) 1×1 conv → also called bottleneck layer.

(*) From 120M to 12M just by using 1×1 conv → Reduce the no. of channels / Increase channels (Modification).

So, —

(1) Modify height, width → Padding, striding, pooling

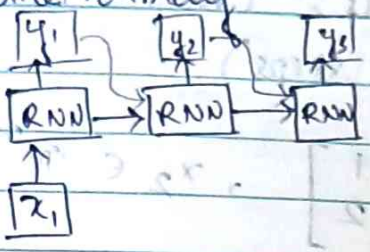
(2) Modify channels → No. of filters, 1×1 conv.

apsara

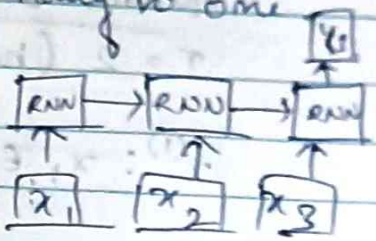
Date 16 09 25

(a) One-to-one

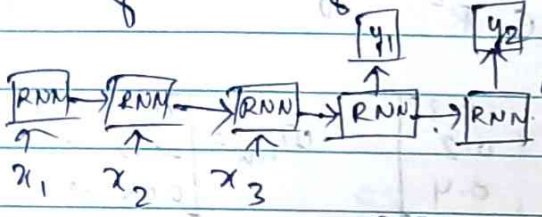
(b) One to many.



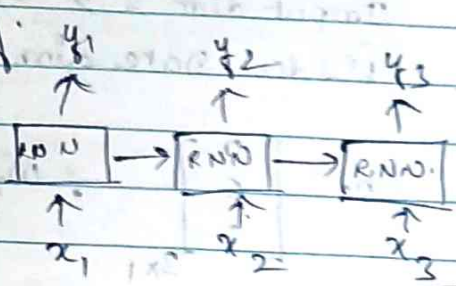
(c) Many to one



(d) many to many.

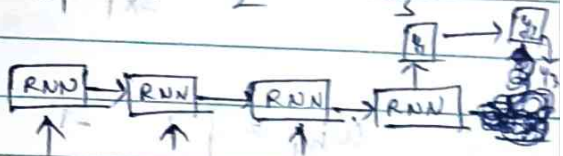


video captioning.



Translation [eg. English to French]

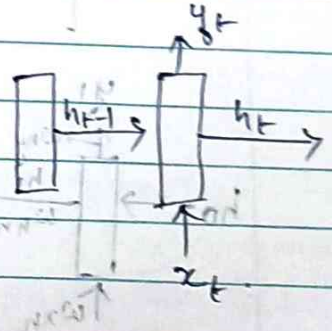
- to see the entire sentence & then give the output.



$$h_t = f_w(h_{t-1}, x_t)$$

$$= \tanh(w_{hh} h_{t-1} + w_{xh} x_t)$$

Note: No bias term.



$$\underset{h_t\ (4\times1)}{\begin{bmatrix}0\\0\\0\\0\end{bmatrix}} = \tanh\left(\underset{w_{hh}\ 4\times4}{\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}} \underset{h_{t-1}\ 4\times1}{\begin{bmatrix}0\\0\\0\\0\end{bmatrix}} + \underset{w_{xh}\ 4\times3}{\begin{bmatrix}0&0&0\\0&0&0\\0&0&0\\0&0&0\end{bmatrix}} \underset{x_t\ 3\times1}{\begin{bmatrix}0\\0\\0\end{bmatrix}} \right)$$

→ Weight sharing at every time stamp.

$$\underset{y_t\ 2\times1}{\begin{bmatrix}0\\0\end{bmatrix}} = \underset{2\times4}{\begin{bmatrix}0&0&0&0\\0&0&0&0\end{bmatrix}} \underset{h_t = 4\times1}{\begin{bmatrix}0\\0\\0\\0\end{bmatrix}}$$

Date 17.09.25

Numerical example :- (vanilla RNN)

T = 2 (two time steps)

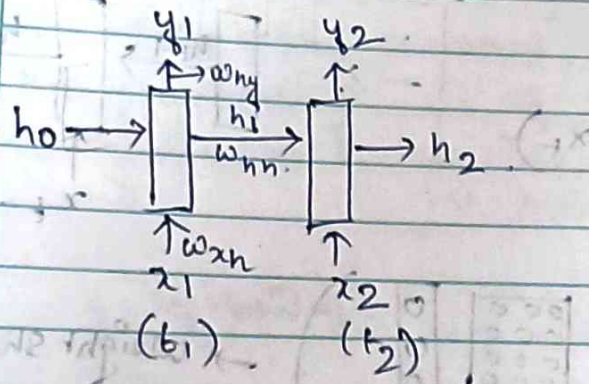
$x^{(1)}$: $x_1 \in \mathbb{R}^2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $x_2 \in \mathbb{R}^2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

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

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

$h_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times1}$, $W_{xh} = \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}_{3\times2}$, $W_{hh} = \begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix}_{3\times3}$

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



$h_1 = \tanh\left(W_{hh} \times h_0 + W_{xh} x_1\right)$

$3\times1 \quad 3\times3 \;\; 3\times1 \quad 3\times2 \;\; 2\times1$ (3×1 after tanh)

$y_1 = W_{hy} \times h_1$

$2\times1 \quad 2\times3 \quad 3\times1$

For t = 1

$W_{hh} \times h_0 + W_{xh} x_1$

$$\begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$2\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix} = \begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix}$$

$$h_1 = \tanh\left(\begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix}\right) = \begin{bmatrix} -0.1 \\ 0.83 \\ 0.716 \end{bmatrix}$$

apsara

Date: ___.___.____

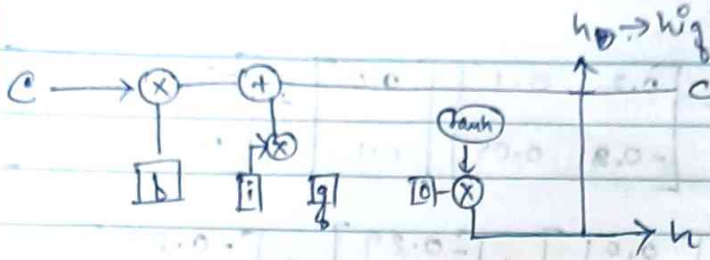
$$y_1 = \begin{bmatrix} 1 & -1 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix} \begin{bmatrix} -0.1 \\ 0.83 \\ 0.716 \end{bmatrix} = \begin{bmatrix} -0.572 \\ 0.007 \end{bmatrix}$$

For $t = 2$

$$w_{hh} \times h_1 + w_{xh} \times x_2$$

$$= \begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} -0.1 \\ 0.83 \\ 0.716 \end{bmatrix} + \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.322 \\ 0.4122 \end{bmatrix} +$$



Numerical example:-

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

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

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

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

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

$W_{xo} = \begin{bmatrix} 0.3 & 0.25 \\ -0.2 & 0.2 \end{bmatrix}$ $W_{ho} = \begin{bmatrix} 0.15 & 0.05 \\ 0.1 & -0.2 \end{bmatrix}$ $W_{xg} = \begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix}$

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

DO- 1 forward pass for 1 time step.
Find - h_t & C_t values.

$$W_{hi} h_{t-1} + W_{xi} x_t = \begin{bmatrix} 0.1 & 0.2 \\ -0.2 & 0.05 \end{bmatrix}_{2\times2} \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}_{2\times1} + \begin{bmatrix} 0.5 & -0.3 \\ 0.4 & 0.1 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.02 \\ 0.005 \end{bmatrix} + \begin{bmatrix} 0.28 \\ 0.19 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.195 \end{bmatrix}$$

$$i_t = \sigma \begin{bmatrix} 0.3 \\ 0.195 \end{bmatrix} = \begin{bmatrix} 0.574 \\ 0.548 \end{bmatrix}$$

$$W_{ho} h_{t-1} + W_{xo} x_t = \begin{bmatrix} 0.15 & 0.05 \\ 0.1 & -0.2 \end{bmatrix} \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix} + \begin{bmatrix} 0.3 & 0.25 \\ -0.2 & 0.2 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.005 \\ -0.02 \end{bmatrix} + \begin{bmatrix} 0.125 \\ -0.12 \end{bmatrix} = \begin{bmatrix} 0.13 \\ -0.14 \end{bmatrix}$$

$$o_t = \sigma \begin{bmatrix} 0.13 \\ -0.14 \end{bmatrix} = \begin{bmatrix} 0.532 \\ 0.468 \end{bmatrix}$$

$\frac{e^x - e^{-x}}{e^x + e^{-x}}$ $\frac{1.323 - 0.7457}{1.323 + 0.7457}$ $\frac{0.5673}{2.0787}$

$\frac{0.2708}{2.0182}$

apsara

Date:

- $w_{hg} h_{t-1} + w_{xg} x_t = \begin{bmatrix} 0.2 & 0.1 \\ -0.21 & 0.05 \end{bmatrix} \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix} + \begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$

$0.9417 - 1.0618$

-0.1201 / 2.0035

-1.030

$$= \begin{bmatrix} 0.01 \\ 0.005 \end{bmatrix} + \begin{bmatrix} -0.29 \\ 0.13 \end{bmatrix} = \begin{bmatrix} -0.28 \\ 0.135 \end{bmatrix}$$

0.9704

$g_t = \tanh \begin{bmatrix} -0.28 \\ 0.135 \end{bmatrix} = \begin{bmatrix} 0.272 \\ 0.134 \end{bmatrix}$

-0.05506 / 2.0004

$w_{hf} h_{t-1} + w_{xf} x_t = \begin{bmatrix} 0.05 & -0.1 \\ 0.2 & 0.1 \end{bmatrix} \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix} + \begin{bmatrix} -0.4 & 0.2 \\ 0.3 & 0.3 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}$

$$= \begin{bmatrix} -0.01 \\ 0.01 \end{bmatrix} + \begin{bmatrix} -0.24 \\ 0.12 \end{bmatrix} = \begin{bmatrix} -0.25 \\ 0.13 \end{bmatrix}$$

$f_t = \sigma \begin{bmatrix} -0.25 \\ 0.13 \end{bmatrix} = \begin{bmatrix} 0.44 \\ 0.53 \end{bmatrix}$

$c_t = \begin{bmatrix} 0.44 \\ 0.53 \end{bmatrix} \odot \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix} + \begin{bmatrix} 0.57 \\ 0.55 \end{bmatrix} \odot \begin{bmatrix} -0.27 \\ 0.13 \end{bmatrix}$

$$= \begin{bmatrix} 0.0888 \\ -0.106 \end{bmatrix} + \begin{bmatrix} -0.1539 \\ 0.0715 \end{bmatrix} = \begin{bmatrix} -0.06 \\ -0.03 \end{bmatrix}$$

$h_t = \begin{bmatrix} 0.53 \\ 0.46 \end{bmatrix} \odot \tanh \begin{bmatrix} -0.06 \\ -0.03 \end{bmatrix} = \begin{bmatrix} 0.53 \\ 0.46 \end{bmatrix} \odot \begin{bmatrix} -0.0599 \\ -0.0297 \end{bmatrix}$

$$= \begin{bmatrix} -0.03 \\ -0.01 \end{bmatrix}$$

Page No.:

apsara

Date: 30 09 25

$x \in \mathbb{R}^2$, $x = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ $\quad W_e \rightarrow 1 \times 2 = [0.5 \quad -1.0]_{1\times 2}$

$h \in \mathbb{R}^1$ $\quad W_d \rightarrow 2 \times 1 = \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix}_{2\times 1}$

↳ context vector

MSE loss $\rightarrow \frac{1}{2} \| \hat{x} - x \|_2^2$

Encoder part

$h = W_e \cdot x = [0.5 \quad -1.0] \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 1$

Decoder part

$\hat{x} = W_d \cdot h = \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix} * 1 = \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix}$

$\text{loss} = \frac{1}{2} \sqrt{(1-2)^2 + (0-0.5)^2}$

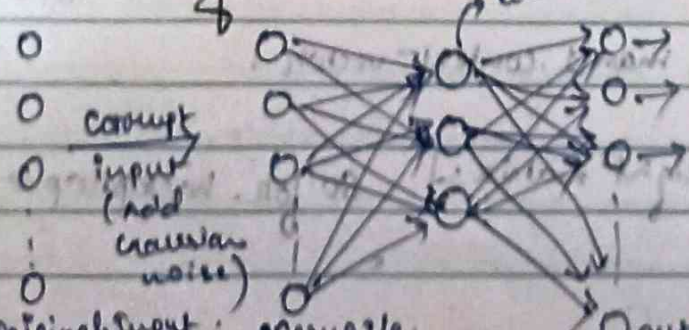
$= \frac{1}{2} \sqrt{1 + 0.25}$

$= \frac{1.118}{2} = 0.559.$

Deep Autoencoders →

Add. Gaussian noise

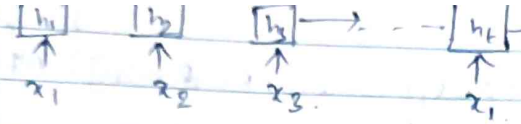
Denoising Autoencoders.



loss = $\tilde{x}$ − original O/P.

learns to remove the noise.

Page No. .....



→ Hidden states of Encoder Network.

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

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

score = similarity / Alignment score.

eg

$h_1 = [1, 0, 1]$

$h_2 = [0, 1, 1]$

$h_3 = [1, 1, 0]$

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

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

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

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

Score func = dot product

$$\alpha_{t,1} = \frac{\exp(2)}{\exp(2) + \exp(1) + \exp(1)} = \frac{7.39}{12.826} = 0.57$$

$$\alpha_{t,2} = \frac{\exp(1)}{12.826} = 0.2$$

$$\alpha_{t,3} = \frac{\exp(1)}{12.826} = 0.2$$

$$c_t = 0.57[1, 0, 1] + 0.2[0, 1, 1] + 0.2[1, 1, 0]$$

$$= [0.57, 0, 0.57] + [0, 0.2, 0.2] + [0.2, 0.2, 0] = [0.77, 0.4, 0.77]$$

3.0629 1.577 · 0.943 / 2.211
0.634

apsara
Date 98.10.25

Numerical example :- Compute z_1 and z_2

Sentence = "Playing Outside"

For playing :-

$q_1 = [0.212 \quad 0.04 \quad 0.63 \quad 0.36]$.

$k_1 = [0.31 \quad 0.84 \quad 0.963 \quad 0.57]$.

$v_1 = [0.36 \quad 0.83 \quad 0.1 \quad 0.36]$.

For outside :-

$q_2 = [0.1 \quad 0.14 \quad 0.86 \quad 0.77]$

$k_2 = [0.45 \quad 0.94 \quad 0.73 \quad 0.58]$

$v_2 = [0.31 \quad 0.36 \quad 0.19 \quad 0.72]$

Note :- use scaled dot product attention.

For playing

$q_1 \cdot k_1 = [0.212 \quad 0.04 \quad 0.63 \quad 0.836] \odot [0.31 \quad 0.84 \quad 0.963 \quad 0.57]$

$= 0.06572 + 0.0336 + 0.60669 + 0.2052$

$= 0.91151$.

Now divide by $\sqrt{4}$ [dimension of k_1].

$\frac{q_1 \cdot k_1}{\sqrt{4}} = \frac{0.91151}{2} = 0.4557$.

~~Softmax (0.4557) = 0.426~~

Now ~~multiply with~~

~~0.926~~

$q_1 \cdot k_2 = [0.212 \quad 0.04 \quad 0.63 \quad 0.36] \odot [0.45 \quad 0.94 \quad 0.73 \quad 0.58]$

$= 0.0954 + 0.0376 + 0.4599 + 0.2088$

$= 0.8017$.

Now divide by $\sqrt{4}$

$\frac{q_1 \cdot k_2}{\sqrt{4}} = \frac{0.8017}{2} = 0.40085$.

Softmax (0.4557, 0.40085) = 0.513, 0.486.

$0.513 * v_1$

$= 0.513 [0.36 \quad 0.83 \quad 0.1 \quad 0.36]$

$= [0.185 \quad 0.426 \quad 0.0513 \quad 0.185]$

$0.486 * [0.31 \quad 0.36 \quad 0.19 \quad 0.72]$

$= [0.150 \quad 0.1746 \quad 0.0921 \quad 0.349]$.

$z_1 = [0.185 \quad 0.426 \quad 0.0513 \quad 0.185] + [0.150 \quad 0.1746 \quad 0.0921 \quad 0.349]$

$= [0.33 \quad 0.605 \quad 0.143 \quad 0.534]$.

3.9961

$v_2 \cdot k_1$

$= [0.1 \quad 0.14 \quad 0.86 \quad 0.77] \odot [0.31 \quad 0.84 \quad 0.963 \quad 0.57]$

$= [0.031 + 0.1176 + 0.828 + 0.4389]$

$= 1.4155$

Now,

$$\frac{v_2 \cdot k_1}{\sqrt{4}} = \frac{1.4155}{2} = 0.707$$

$v_2 \cdot k_2$

$[0.1 \quad 0.14 \quad 0.86 \quad 0.77] \odot [0.45 \quad 0.94 \quad 0.73 \quad 0.58]$

$= [0.045 + 0.1316 + 0.6278 + 0.4466] = 1.251$

Now,

$$\frac{v_2 \cdot k_2}{\sqrt{4}} = \frac{1.251}{2} = 0.6255$$

Softmax $(0.707, 0.6255) = 0.52, 0.48$

$0.52 * v_2$

$= 0.52 [0.31 \quad 0.36 \quad 0.19 \quad 0.72]$

$= [0.1612 \quad 0.1872 \quad 0.0988 \quad 0.3744]$

$0.48 * v_2$

$= 0.48 [0.31 \quad 0.36 \quad 0.19 \quad 0.72]$

$[0.1488 \quad 0.1728 \quad 0.0912 \quad 0.3456]$

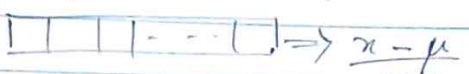
$z_2 = [0.1612 \quad 0.1872 \quad 0.0988 \quad 0.3744] + [0.1488 \quad 0.1728 \quad 0.0912 \quad 0.3456]$

$= [0.31 \quad 0.36 \quad 0.19 \quad 0.72]$

apsara

Date: 13.10.25

Layer Normalization.

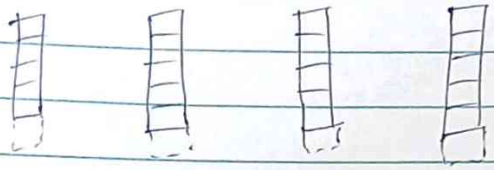


$\Rightarrow \frac{x-\mu}{\sigma}$

Positional encoding :-

Key components of Transformer architecture.

0 1 2 3 4



Sinusoidal & cosine func^{n}.

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

$$PE_{pos,2i+1} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{emb}}}}\right).$$

positional embedding vector.

Even pos → use sine

Odd pos → use cosine.

eg: I am a robot.

P_{E}, $d = 4 \Rightarrow R^4$

Positional Embedding

I	→	0	→	P_{00}	P_{01}	P_{02}	P_{03}	→	0 | 1 | 0 | 1
am	→	1	→	P_{10}	P_{11}	P_{12}	P_{13}	→	0.84 | 0.54 | 0.10 | 1
a	→	2							
Robot	→	3	→	P_{30}	P_{31}	P_{32}	P_{33}		

$$P_{00} = \sin\left(\frac{0}{10000^{0/4}}\right) = \sin(0) = 0$$

$$P_{01} = \cos\left(\frac{0}{10000^{0/4}}\right) = \cos(0) = 1$$

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

$$P_{03} = \cos\left(\frac{0}{10000^{2/4}}\right) = \cos(0) = 1$$

$$P_{10} = \sin\left(\frac{1}{10000^{0/4}}\right) = \sin(1) = 0.8414$$

$$P_{11} = \cos\left(\frac{1}{10000^{0/4}}\right) = \cos(1) = 0.54$$

$$P_{12} = \sin\left(\frac{1}{10000^{2/4}}\right) = 0.10$$

$$P_{13} = \cos\left(\frac{1}{10000^{2/4}}\right) = 0.999 = 1$$

Page No.: