softmax

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \Rightarrow \begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix}$$

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

$$J = \begin{bmatrix} \frac{\partial S_1}{\partial x_1} & \frac{\partial S_1}{\partial x_2} & \frac{\partial S_1}{\partial x_3} \\ \frac{\partial S_2}{\partial x_1} & \frac{\partial S_2}{\partial x_2} & \frac{\partial S_2}{\partial x_3} \\ \frac{\partial S_3}{\partial x_1} & \frac{\partial S_3}{\partial x_2} & \frac{\partial S_3}{\partial x_3} \end{bmatrix}$$

(Jacobian matrix)

For diagonal elements $(i=j)$: $\frac{\partial S_1}{\partial x_1} = S_1(1-S_1)$

For non-diagonal elements $(i \neq j)$: $\frac{\partial S_1}{\partial x_2} = -S_i S_j$

Q. $x = \begin{bmatrix} 2.0 \\ 1.0 \\ 0.1 \end{bmatrix}$

compute softmax o/p
compute the Jacobian

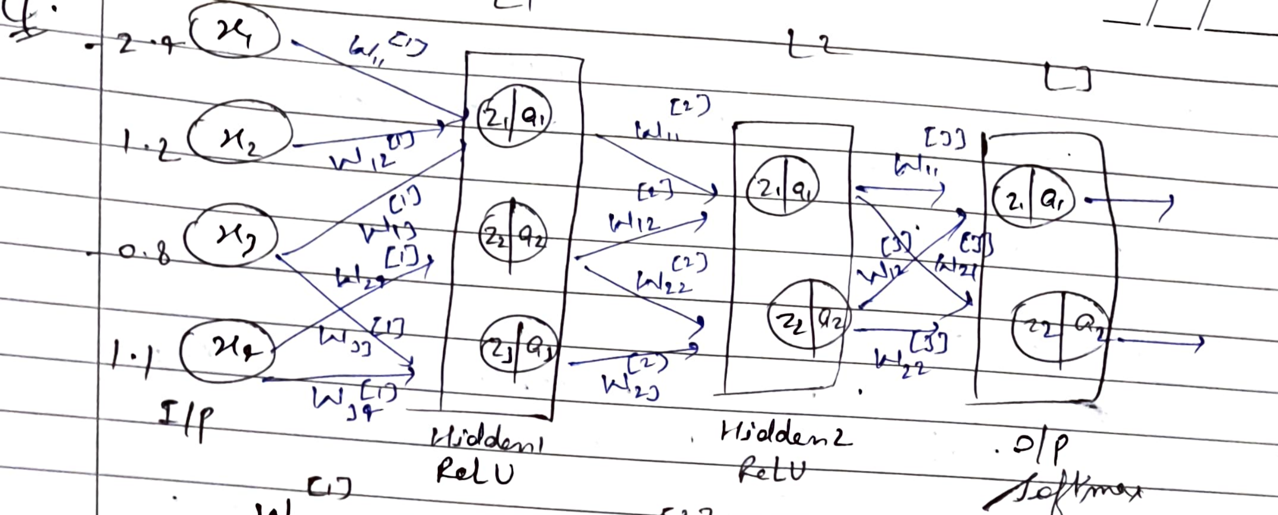
$$S_1 = \frac{7.4}{7.4 + 2.7 + 1.1} = \frac{7.4}{11.2} = 0.66$$

$$S_2 = \frac{2.7}{11.2} = 0.24$$

$$S_3 = \frac{1.1}{11.2} = 0.1$$

$$J = \begin{bmatrix} 0.2244 & -0.1584 & -0.066 \\ -0.1584 & 0.1824 & -0.024 \\ 0.066 & -0.024 & 0.09 \end{bmatrix}_{3\times3}$$

Q.



$W^{[1]} = 0.1$, $W^{[2]} = 0.001$, $W^{[3]} = 0.01$

For layer 1,

$$z_1 = W_{11}^{[1]} x_1 + W_{12}^{[1]} x_2 + W_{13}^{[1]} x_3 + W_{14}^{[1]} x_4$$
$$= 0.1(-2.4) + 0.1(1.2) + (0.1)(-0.8) + (0.1)(1.1)$$
$$= -0.24 + 0.12 - 0.08 + 0.11 = -0.09.$$

ReLU

$a_1 = 0$

$z_2 = -0.09$

$a_2 = 0$

$z_3 = -0.09$

$a_3 = 0$

For layer 2

$$z_1 = W_{11}^{[2]} a_1 + W_{12}^{[2]} a_2 + W_{13}^{[2]} a_3$$
$$= 0.001 \times 0 + 0.001 \times 0 + 0.001 \times 0 = 0$$

ReLU

$a_1 = 0$

$z_2 = 0$

$a_2 = 0$

For layer 3,

$$z_1 = W_{11}^{[3]} a_1 + W_{12}^{[3]} a_2 = 0 \times 0.01 + 0 \times 0.01 = 0$$

Softmax

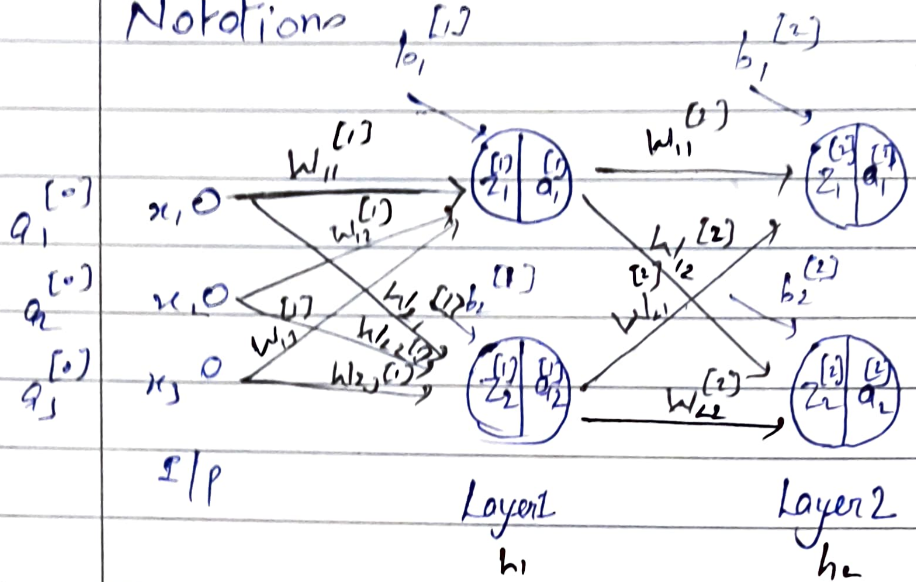
$a_1 = 0.5$ because $\frac{e^0}{e^0 + e^0} = \frac{1}{2} = 0.5$

$z_2 = 0$ Hence $\hat{y}_1 = 0.5$ and $\hat{y}_2 = 0.5$

$a_2 = 0.5$

Feed Forward Neural Network

Notations



$W^{[l]}$ → connections – matrix

W rows – no. of neurons in l^{th} layer

cols → no. of neurons in $(l-1)^{th}$ layer

$b^{[l]}$ – bias

$z^{[l]}$ – $w($ $) + b$

$a^{[l]}$ – $g(z)$ → elementwise operation

$a^{[0]}$ – x

~~W rows~~

In this case,

$$W^{[1]} = \begin{bmatrix} W_{11}^{[1]} & W_{12}^{[1]} & W_{13}^{[1]} \\ W_{21}^{[1]} & W_{22}^{[1]} & W_{23}^{[1]} \end{bmatrix} \qquad b = \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \end{bmatrix} \qquad z^{[1]} = \begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \end{bmatrix}$$

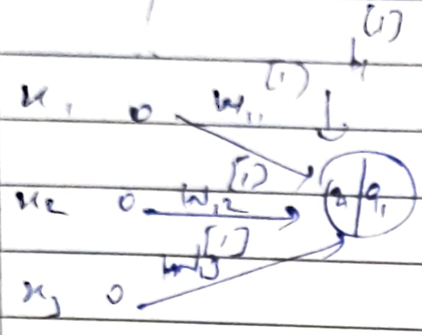
$$z_1^{[1]} = W_{11}^{[1]} a_1^{[0]} + W_{12}^{[1]} a_2^{[0]} + W_{13}^{[1]} a_3^{[0]} + b_1^{[1]}$$

$$z_2^{[1]} = W_{21}^{[1]} a_1^{[0]} + W_{22}^{[1]} a_2^{[0]} + W_{23}^{[1]} a_3^{[0]} + b_2^{[1]}$$

$$z_1^{[2]} = W_{11}^{[2]} a_1^{[1]} + W_{12}^{[2]} a_2^{[1]} + \cancel{W_{13}^{[2]} a_3^{[1]}} + b_1^{[2]}$$

$$z_2^{[2]} = W_{21}^{[2]} a_1^{[1]} + W_{22}^{[2]} a_2^{[1]} + \cancel{W_{23}^{[2]} a_3^{[1]}} + b_2^{[2]}$$

Perceptron - single layer NN and with only one neuron



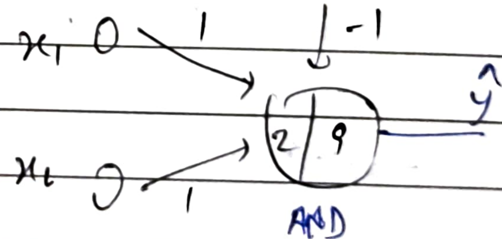
$$z^{[1]} = W_{11}^{(1)} x_1 + W_{12}^{[1]} x_2 + W_{13}^{[1]} x_3 + b_1^{[1]}$$

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

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

AND

	x_1	x_2	y
1)	0	0	0
2)	0	1	0
3)	1	0	0
4)	1	1	1



1) $0 + 0 - 1 = -1 \longrightarrow 0$ ~~$= -1 \quad 0 \longrightarrow 0$~~

2) $0 + 1 - 1 = 0 \longrightarrow 0$

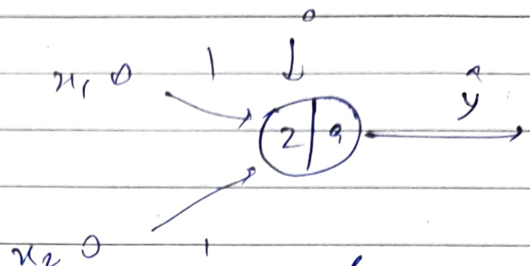
3) $1 + 0 - 1 = 0 \longrightarrow 0$

4) $1 + 1 - 1 = 1 \longrightarrow 1$

1x2 2x2 1x2

__/__/____

For OR gate



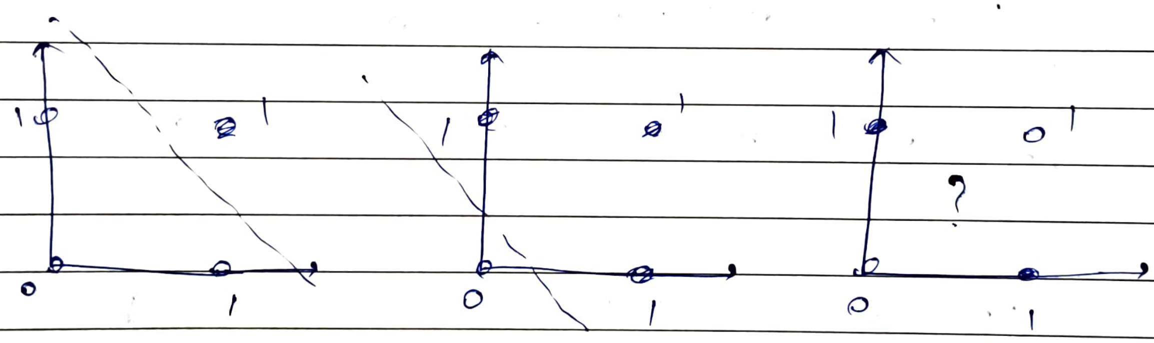
OR

1) $0 + 0 + 0 = 0 \to 0$
2) $1 + 0 + 0 = 1 \to 1$
3) $0 + 1 + 0 = 1 \to 1$
4) $1 + 1 + 0 = 2 \to 1$

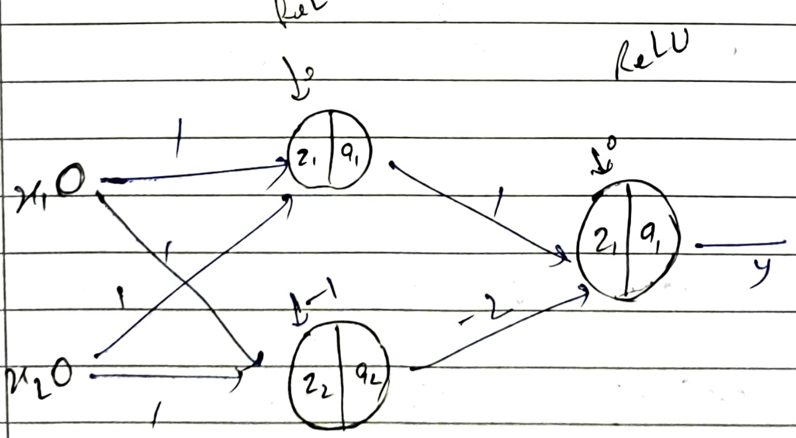
For XOR can't be modelled using perceptron

0	0	0
1	0	1
0	1	1
1	1	0

$W_{11} = W_{12} = 1$ and bias.



AND OR XOR



1)	0	0	0
2)	0	1	1
3)	1	0	1
4)	1	1	0

1) ~~$0 \times 1 + 0 \times 1 + 0 = 0 \to 0 \times 1 + 0 \times -2 = 0 + 0 \to 0$~~
2) ~~$0 \times 1 + 1 \times 1 + 0$~~ ~~$1 \times 0 + -2 \times 1 = 0 - 2 = -2$~~
3) ~~$1 \times 1 + 0 \times -2$~~
4) ~~$1 + 0$~~ -2

1) L1
$z_1 = 0\times1 + 0\times1 + 0 = 0$ $\quad a_1 = 0$

$z_2 = 0\times1 + 0\times1 - 1 = -1$ $\quad a_2 = 0$

L2
$z_1 = 0\times1 + 0\times-2 + 0 = 0$

$a_1 = 0 \longrightarrow \hat{y}$

2) L1
$z_1 = 1\times1 + 0\times1 + 0 = 1$ $\quad a_1 = 1$

$z_2 = 1\times1 + 0\times1 - 1 = 0$ $\quad a_2 = 0$

L2
$z_1 = 1\times1 + 0\times-2 + 0 = 1$ $\quad a_1 = 1 \rightarrow \hat{y}$

3) L1
$z_1 = 0\times1 + 1\times1 + 0 = 1$ $\quad a_1 = 1$

$z_2 = 0\times1 + 1\times1 - 1 = 0$ $\quad a_2 = 0$

L2
$z_1 = 1\times1 + 0\times-2 + 0 = 1$ $\quad a_1 = 1 \rightarrow \hat{y}$

4) L1
$z_1 = 1\times1 + 1\times1 + 0 = 2$ $\quad a_1 = 2$

$z_2 = 1\times1 + 1\times1 - 1 = 1$ $\quad a_2 = 1$

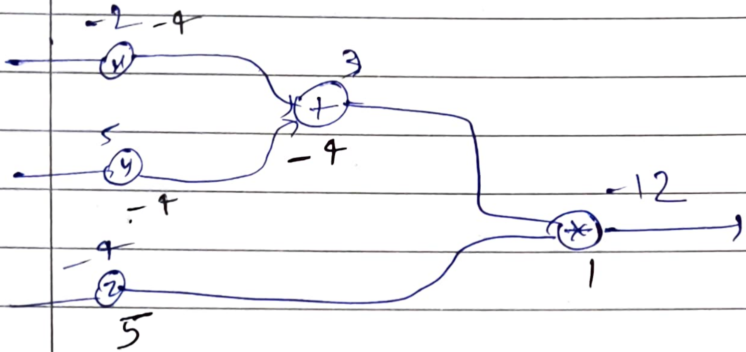
$z_1 = 2\times1 + 1\times-2 + 0 = 0$ $\quad \hat{y} = 0$

29/7/25 SSR

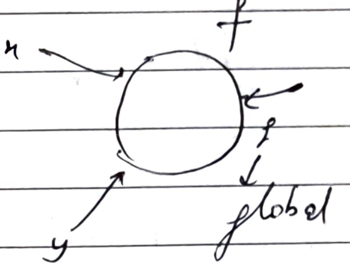
Computational graph

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

$\Rightarrow q = (x+y)$



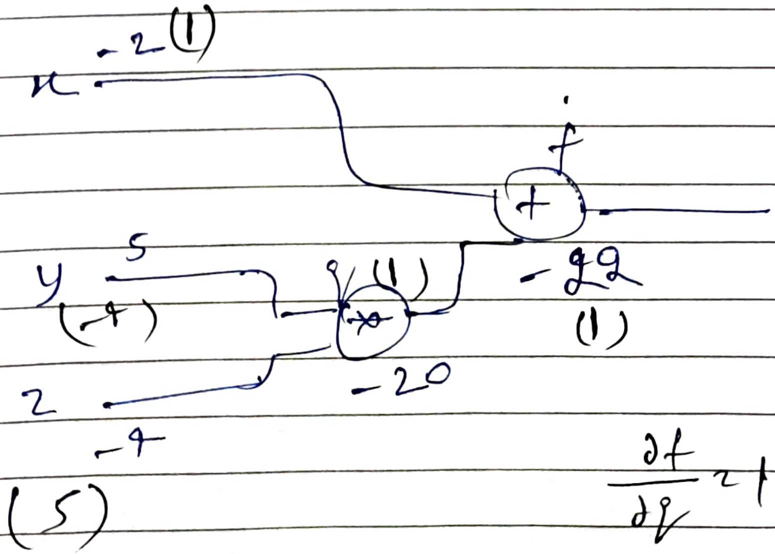
local gradient (*)



Q. $f(x, y, z) = x + yz \Rightarrow f = x + q$; $q = yz$

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

1) Identify additional function
1.1) compute local derivatives
2) Draw a computational graph
3) Perform forward pass
4) Perform backward pass starting from the end of the circuit



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

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

$\frac{\partial f}{\partial y} = 2$

$\frac{\partial f}{\partial q} = 1 \quad \frac{\partial f}{\partial z} = y$

$$\frac{\partial f}{\partial z} = \text{local gradient} \times \text{global gradient}$$

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

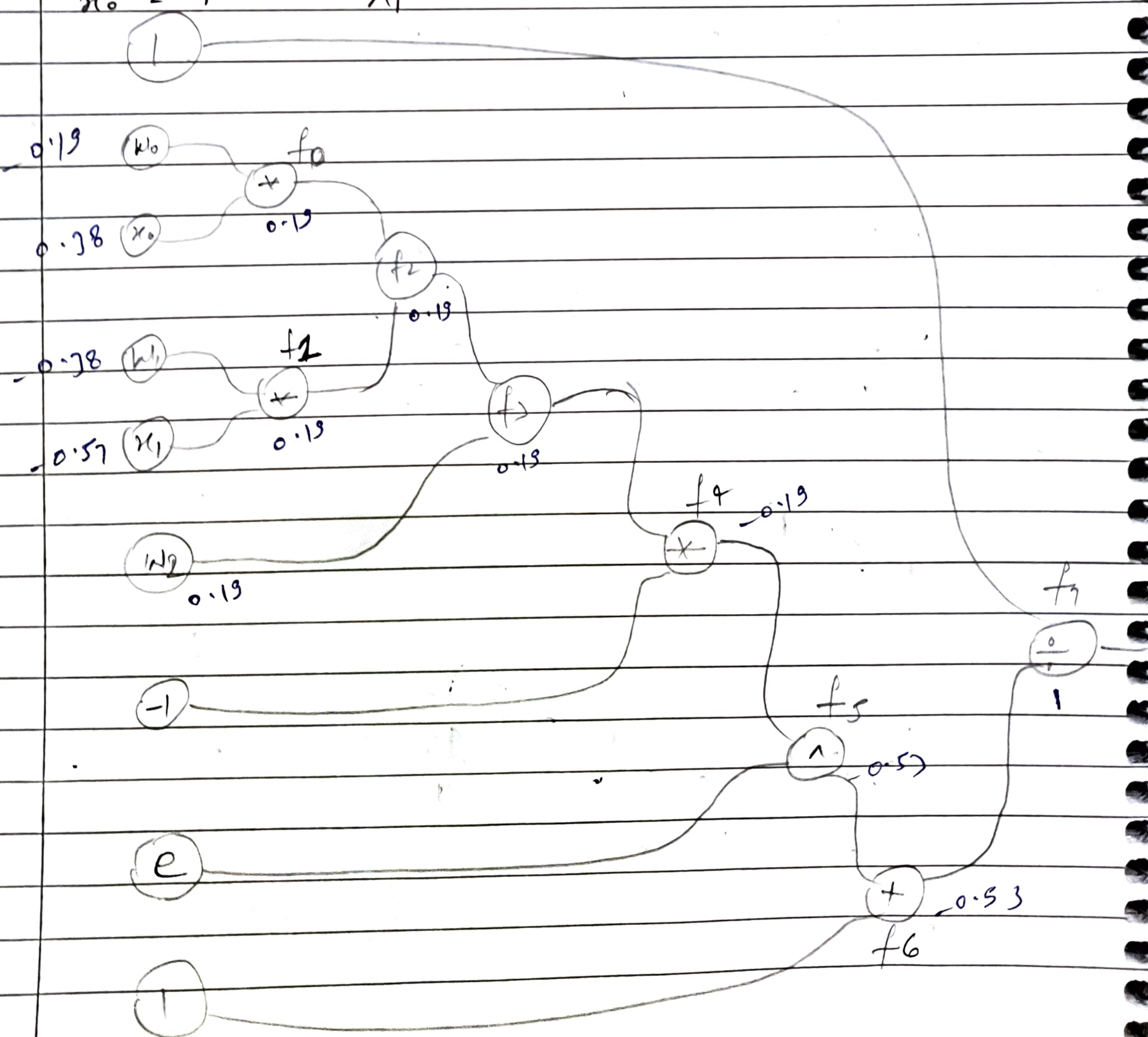
$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial y} \times 1 = -4 \times 1 = -4$$

Q. $$f(w, x) = \frac{1}{1 + e^{-(\underbrace{w_0 x_0}_{f_1} + \underbrace{w_1 x_1}_{f_2} + w_2)}}$$

$W_0 = 2$ $\quad$ $w_1 = -3$ $\quad$ $W_2 = -3$

$x_0 = -1$ $\quad$ $x_1 = -2$

Diag



less chances of missing

30/7/25

Forward pass

$f_0 = -2 = w_0 x_0$

$f_1 = +6 = w_1 x_1$

$f_2 = 4 \quad f_2 + w$

$f_3 = 1 = \cancel{f_1 + f_2 + w_2}$

$f_4 = -f_3 = -1$

$f_5 = e^{f_4} = e^{-1} = 0.37$

$f_6 = 1 + f_5 = 1.37$

$f_7 = \frac{1}{f_6} = \frac{1}{1.37} = 0.73$

Backpropagation

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

$\frac{\partial f_7}{\partial f_6} = -\frac{1}{(1.37)^2} = \frac{1}{1.8769} = -0.53$

$\frac{\partial f_7}{\partial f_5} = \frac{\partial f_7}{\partial f_6} \times \frac{\partial f_6}{\partial f_5} = 0.53 \times 1 = -0.53$

$$\frac{\partial f_7}{\partial f_4} = \frac{\partial f_7}{\partial f_5} \times \frac{\partial f_5}{\partial f_4} = -0.53 \times 0.37 = -0.1961$$

$$\frac{\partial f_7}{\partial f_3} = \frac{\partial f_7}{\partial f_4} \times \frac{\partial f_4}{\partial f_3} = -0.1961 \times (-1) = +0.1961$$

$$\frac{\partial f_7}{\partial f_2} = \frac{\partial f_7}{\partial f_3} \times \left[\frac{\partial f_3}{\partial f_2} + \frac{\partial f_3}{\partial f_1} + \frac{\partial f_3}{\partial w_2}\right]$$

$$= 0.1961 \times \left[1 + 1 + 1\right] = 0.5883$$

$$\frac{\partial f_7}{\partial f_2} = \frac{\partial f_7}{\partial f_3} \times \left(\frac{\partial f_3}{\partial f_2} + \frac{\partial f_3}{\partial w_2}\right) = 0.1961 \times (2) = 0.3922$$

$$\frac{\partial f_7}{\partial f_1} = \frac{\partial f_7}{\partial f_2} \times \left[\frac{\partial f_2}{\partial f_1} + \frac{\partial f_2}{\partial f_0}\right] = 0.39 \times 2 = 0.78\cancel{4}$$

$$\frac{\partial f_7}{\partial f_0} =$$

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

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

$$\frac{\partial f_7}{\partial f_1} = \frac{\partial f_7}{\partial f_2} \times \frac{\partial f_2}{\partial f_1} = 0.19 \times 1 = 0.19$$

$$\frac{\partial f_7}{\partial f_0} = \frac{\partial f_7}{\partial f_2} \times \frac{\partial f_2}{\partial f_0} = 0.19 \times 1 = 0.19$$

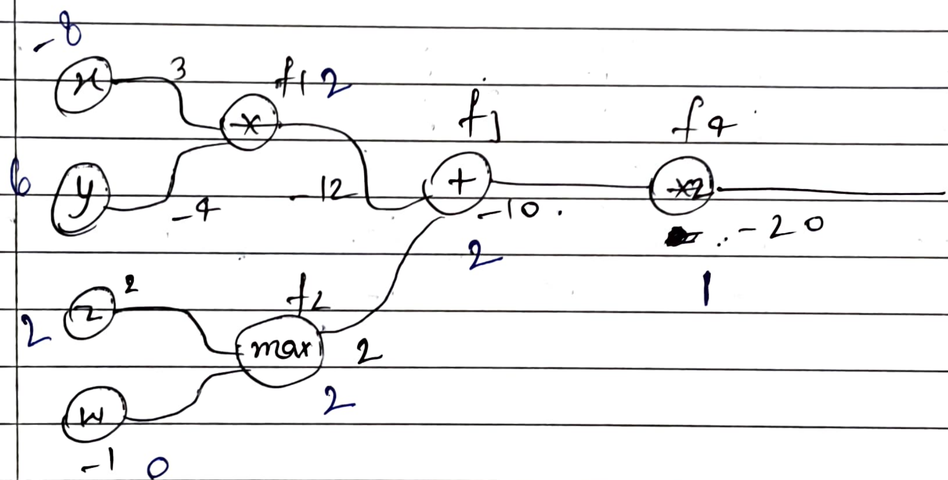
$$\frac{\partial f_7}{\partial w_0} = \frac{\partial f_7}{\partial f_0} \times \frac{\partial f_0}{\partial w_0} = 0.19 \times -1 = -0.19$$

$$\frac{\partial f_7}{\partial x_0} = \frac{\partial f_7}{\partial f_0} \times \frac{\partial f_0}{\partial x_0} = 0.19 \times 2 = 0.38$$

$$\frac{\partial f_7}{\partial w_1} = \frac{\partial f_7}{\partial f_1} \times \frac{\partial f_1}{\partial w_1} = 0.19 \times -2 = -0.38$$

$$\frac{\partial f_7}{\partial x_1} = \frac{\partial f_7}{\partial f_1} \times \frac{\partial f_1}{\partial x_1} = 0.19 \times -3 = -0.57$$

Q.



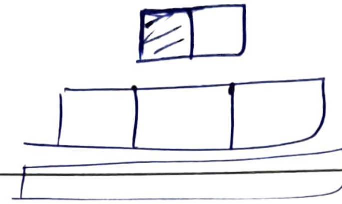
$$f_1 = x \times y = 3 \times (-4) = -12$$

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

$$f_3 = f_1 + f_2 = -10$$

$$f_4 = f_3 \times 2 = -20$$

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



___/___/______

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

$$\frac{\partial f_4}{\partial f_1} = \frac{\partial f_4}{\partial f_3} \times \frac{\partial f_3}{f_1} = 2 \times 1 = 2.$$

$$\frac{\partial f_4}{\partial f_2} = \frac{\partial f_4}{\partial f_3} \times \frac{\partial f_3}{\partial f_2} = 2 \times 1 = 2.$$

$$\frac{\partial f_4}{\partial x} = \frac{\partial f_4}{\partial f_1} \times \frac{\partial f_1}{\partial x} = 2 \times -4 = -8$$

$$\frac{\partial f_4}{\partial y} = \frac{\partial f_4}{\partial f_1} \times \frac{\partial f_1}{\partial y} = 2 \times 3 = 6$$

$$\frac{\partial f_4}{\partial z} = \frac{\partial f_4}{\partial f_2} \times \frac{\partial f_2}{\partial z}$$

$z > w \rightarrow 1$

$z < w \rightarrow 0$

$= 2 \times 1 = 2$

$z = w$ = undefined

$$\frac{\partial f_4}{\partial w} = \frac{\partial f_4}{\partial f_2} \times \frac{\partial f_2}{\partial w}$$

$w > z \rightarrow 1$

$= 2 \times 0 = 0$

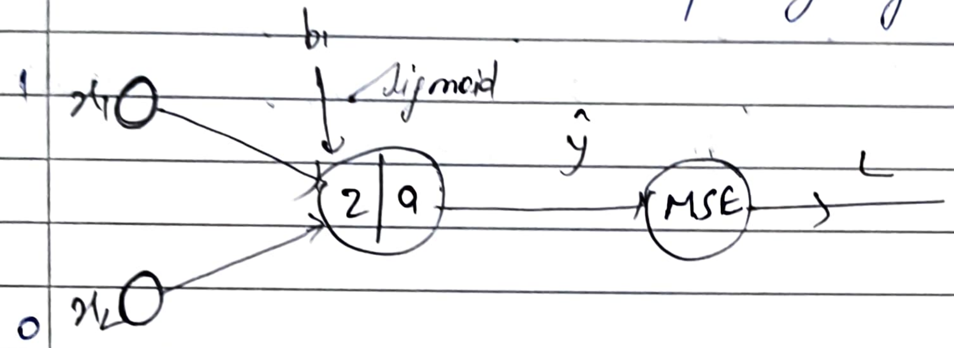
$w < z \rightarrow 0$

$w = z \rightarrow$ undefined

31/7/25 ⊕ sum gate → gradient will get distributed equally to all its inputs

max gate → routes the gradient to the higher i/p path

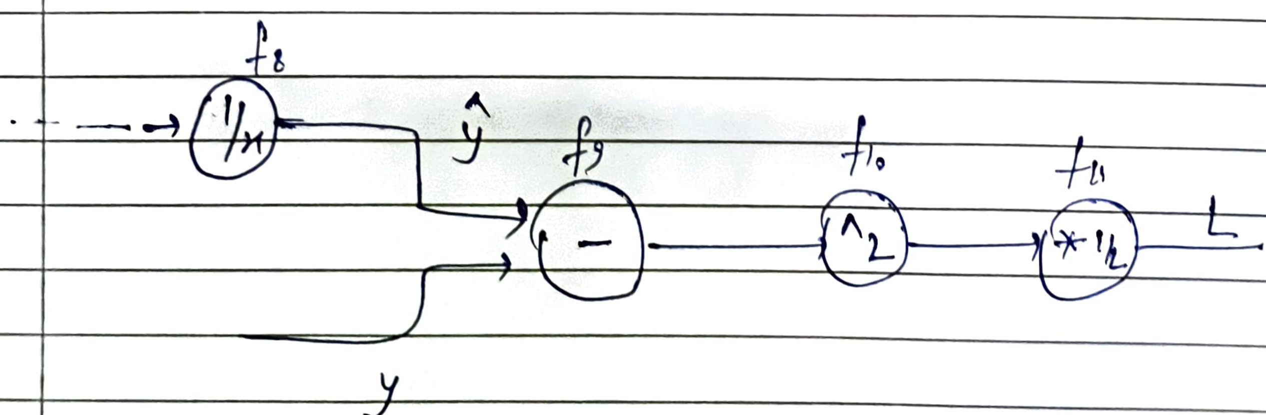
⊛ multiply gate → swaps i/p activation and multiply by global gradient.



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

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

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

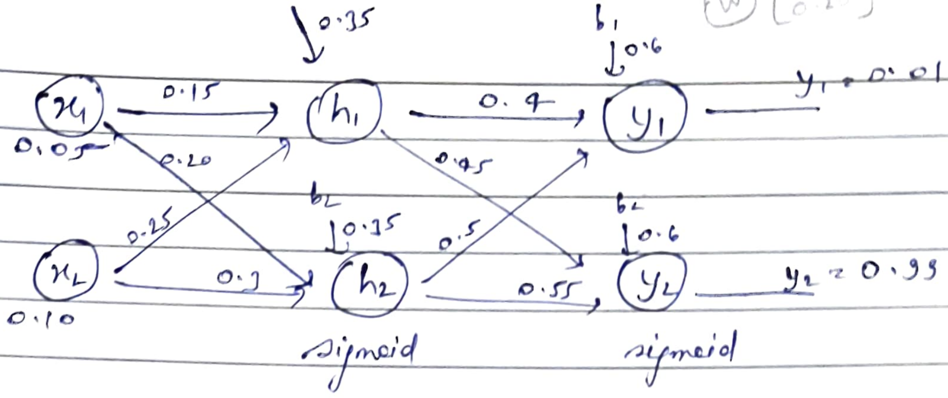


- Can do sigmoid directly as a node as we know the derivative $(\sigma(z))(1 - \sigma(z))$

21

W [0.15 0.20]

__/__/__

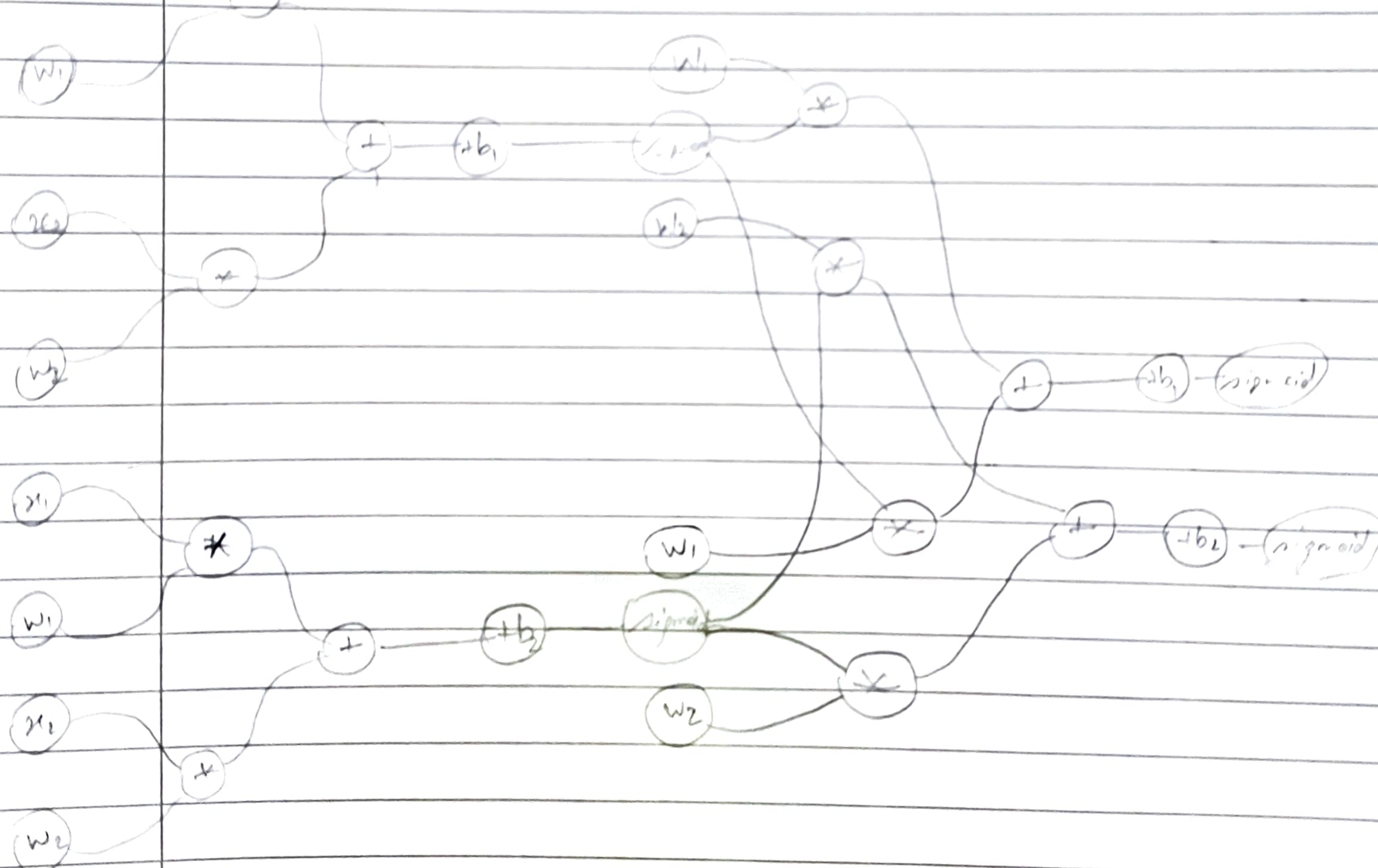


Loss = $\Sigma \frac{1}{2}$ (target − output)^2

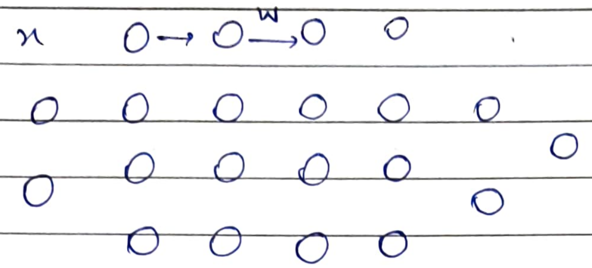
for all neuron in o/p layer

$h_{1(2)} = x_1 w_1 + x_2 w_2 + b_1$

a =



6/8/25



we will do normalization before non linear activation func^n

scale and shift

$$W_{new} = W_{old} - \alpha \frac{\partial L}{\partial W}$$

→ Normalize each layer
→ Apply it to z
→ scale and shift → learnable parameters

[γ] (β)

— $\gamma^{[1]}$ $\beta^{[1]}$

Q. $z = [4, 8, 6, 10]$ — Apply it to every batch

$$\mu = \frac{4+8+6+10}{4} = 7$$

$$\sigma = \frac{1}{4} \sum_{i=1}^{4} (z_i - \mu)^2 = 5$$

$$= \frac{(4-7)^2 + (8-7)^2 + (6-7)^2 + (10-7)^2}{4+\epsilon} = 5$$

can be something very small like 10^{-5} to avoid the undefined situation

$$z = \frac{x-\mu}{\sigma+\epsilon}$$

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

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

γ, β are learnable parameters

7 × 6 $\quad$ 3 × 3 $\quad$ 5 × 4 $\quad$ __/__/__

Convolutional operation

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

$*$

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

$=$

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

convolution operation

input 6 × 6 image

filter or kernel 3 × 3

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

$= 3+0+\cancel{-1}+\cancel{1}+0-8+\cancel{2}+0\cancel{-2} = -5$

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

$= 0+0-2+\cancel{5}+0-9+7+0\cancel{-5} = -4$

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

$= 1+0-7+8+0\cancel{-3}+\cancel{2}+0-1 = 0$

→ Vertical and Horizontal Edge Detection

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

Vertical

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

Horizontal

19/8/25

0	0	0	0	0	0
0	2	8	7	4	6
0	[illegible]	6	9	5	7
0	3	4	8	3	8
0	7	5	3	6	6
0	4	2		[illegible]	3
0	0		0		

$n = 5 \times 5$

$\times$

1	-1
0	-1

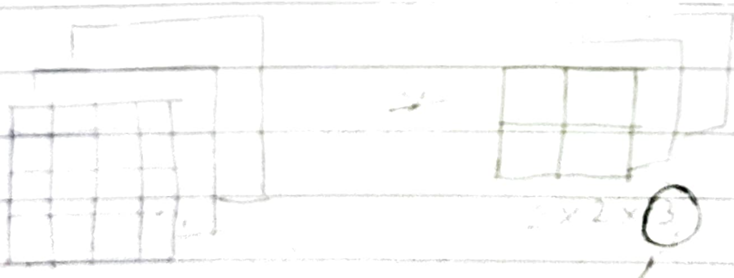
$f = 2 \times 2$

$P = 1 \quad S = 2$

o/p = 4×4

-2	-7	-6
-3	-11	-7
-11	4	-3

convolutions on a 3D volume



$2 \times 2 \times 3$

$4 \times 4 \times 3$ → Has to be equal

Height width channels

19/8/25

10 filters } in one layer of
3 × 3 × 3 } conv net

How many parameters does this layer have?

$W \rightarrow 3 \times 3 \times 3 \times 10 = 270$

$b \rightarrow 10 \rightarrow$ so $270 + 10 = 280$

1) CONV layer
2) Pooling layer — Max Pooling / Avg Pooling
3) Fully connected (FC) layer

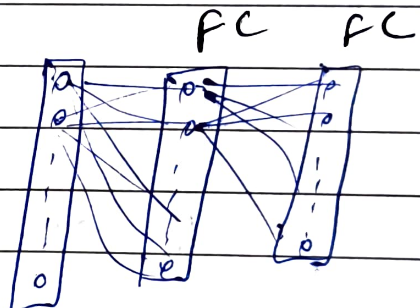
1	3	2	1	3
2	9	1	1	5
1	3	2	3	2
8	3	5	1	0
5	6	1	2	9

$5 \times 5 \times n_c$

9	9	5
9	9	5
8	6	9

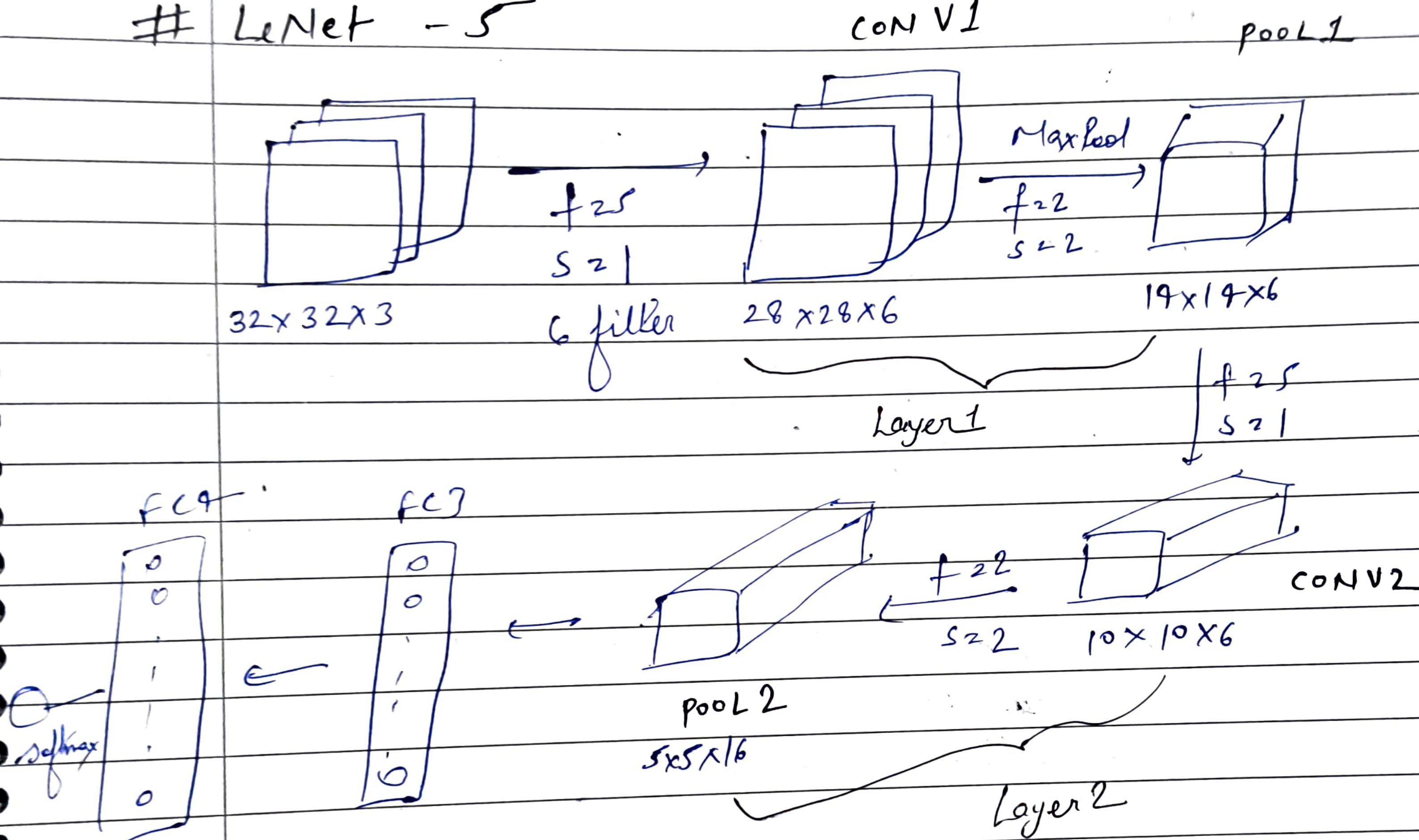
$3 \times 3 \times n_c$

=



Pooling → slide and take the max.
· It can be affected by striding and padding

LeNet -5



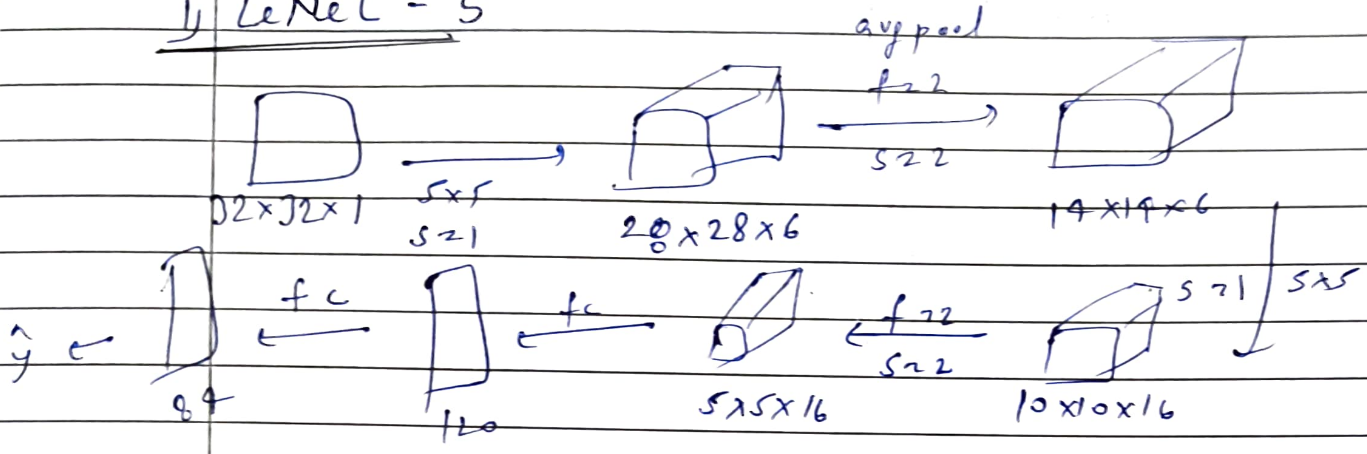
input → CONV1 → POOL1 → CONV2 → POOL2

↓

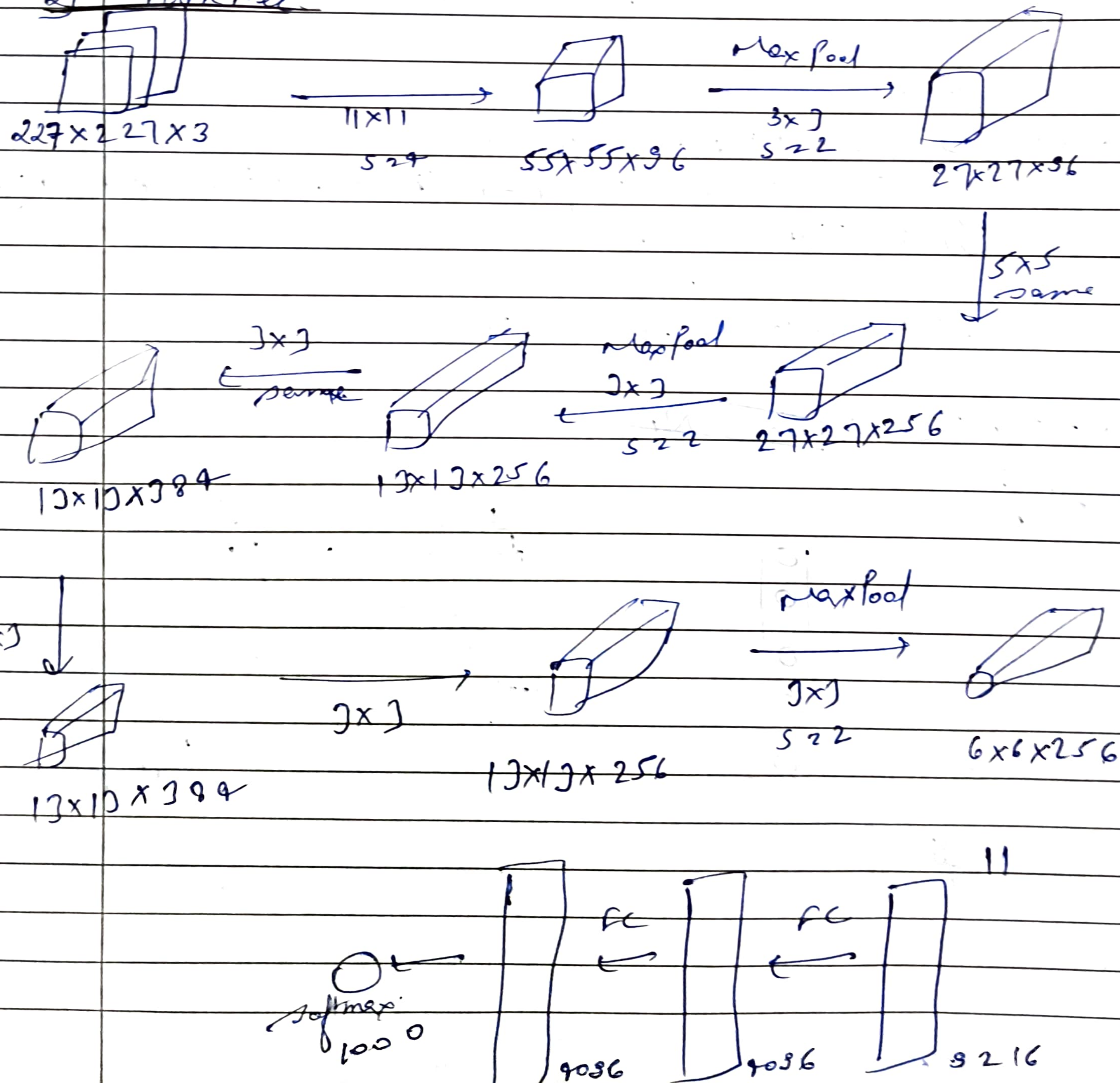
softmax ← FC4 ← FC3

CLASSIC NETWORKS

1) LeNet - 5



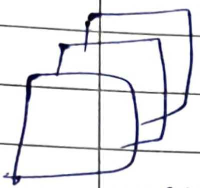
2) AlexNet



3) VGG-16

CONV = 3×3 filter, s = 1, same

MAX-POOL = 2×2, s = 2



$224 \times 224 \times 3 \xrightarrow[\times 2]{[\text{CONV64}]} 224 \times 224 \times 64 \xrightarrow{\text{POOL}} 112 \times 112 \times 64$

$\xrightarrow{[\text{CONV128}]} 112 \times 112 \times 128 \xrightarrow{\text{Pool}} 56 \times 56 \times 128$

$\xrightarrow{} 56 \times 56 \times 256$ ~~28×28×256~~

$\xrightarrow[\times 3]{[\text{CONV256}]} 28 \times 28 \times 256$

$\xrightarrow{\text{Pool}} 28 \times 28 \times 512$

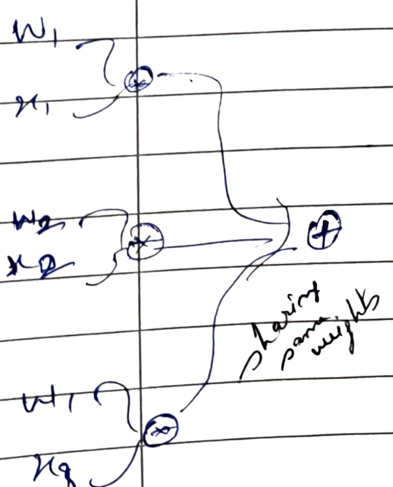
~~[CONV512] ×3~~

$\longrightarrow 14 \times 14 \times 512 \xrightarrow[\times 3]{[\text{CONV512}]} 14 \times 14 \times 512 \xrightarrow{\text{POOL}} 7 \times 7 \times 512$

$\longrightarrow$ 4096 FC $\longrightarrow$ 4096 FC $\longrightarrow$ 1000 softmax

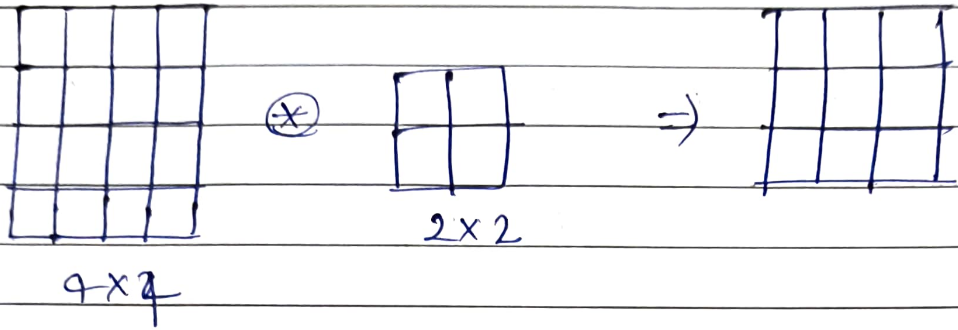
BACKPROP

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



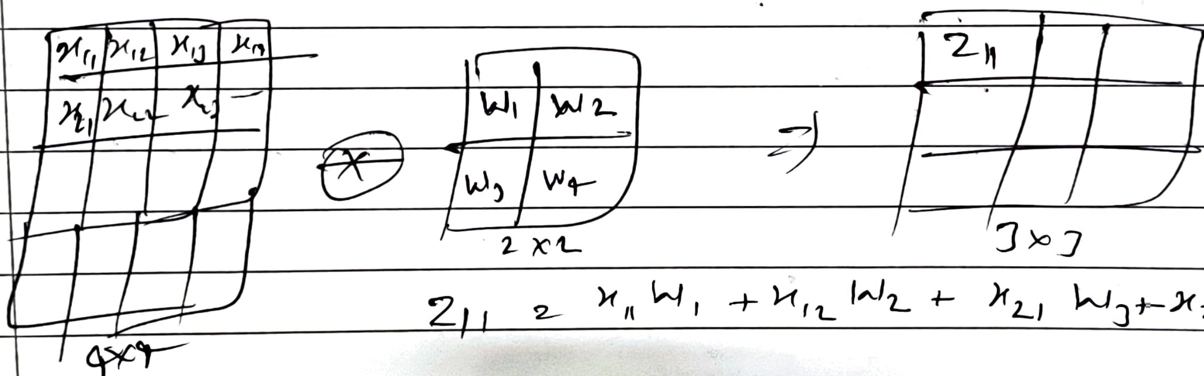
← coming from different paths

sum the gradient of parameter



I/p → CONV1 → POOL1 → CONV2 → POOL2 → CONV3 → POOL3 → FC4 → FC5 → Softmax

$$\frac{\partial L}{\partial W_1}, \frac{\partial L}{\partial W_2}, \frac{\partial L}{\partial W_3}, \frac{\partial L}{\partial W_4}$$



$$z_{11} = x_{11} w_1 + x_{12} w_2 + x_{21} w_3 + x_{22} w_4$$

25/8/25 chr22

- ResNet
- 1x1 conv
- InceptionNet
- CNN / ViT

$$a^{[l+2]} = g\left(W^{[l+2]} \times a^{[l+1]} + b^{[l+2]} + a^{[l]}\right)$$

$$= g\left(z^{[l+2]} + a^{[l]}\right)$$

256 128

$\Downarrow$

$$= g\left(z^{[l+2]} + \boxed{W} a^{[l]}\right)$$

Eg:- $\underset{\cancel{64 \times 1} \; 32\times 1}{a^{[l+2]}} = \text{ReLU}\left(\underset{32\times 1}{z^{[l+2]}} + \underset{\boxed{W}\, 64\times 1}{a^{[l]}}\right)$

__/__/__

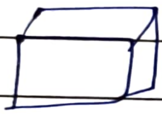
ResNet - 16

- 32

-

1x1 conv

$$\begin{bmatrix}1&2&3\\4&5&6\\7&8&9\end{bmatrix}_{3\times3} * \boxed{2} = \begin{bmatrix}2&4&6\\ & & \\ & & \end{bmatrix}_{3\times3}$$



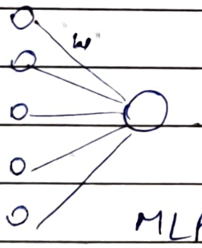
6x6x32

*

1x1x32

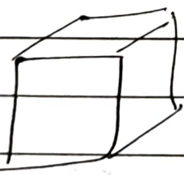
+ ReLU

"Network in a network"



MLP

Eg.

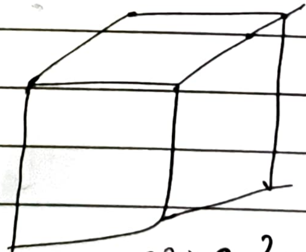


28x28x192

CONV → 5x5x192

#filters = 32

same

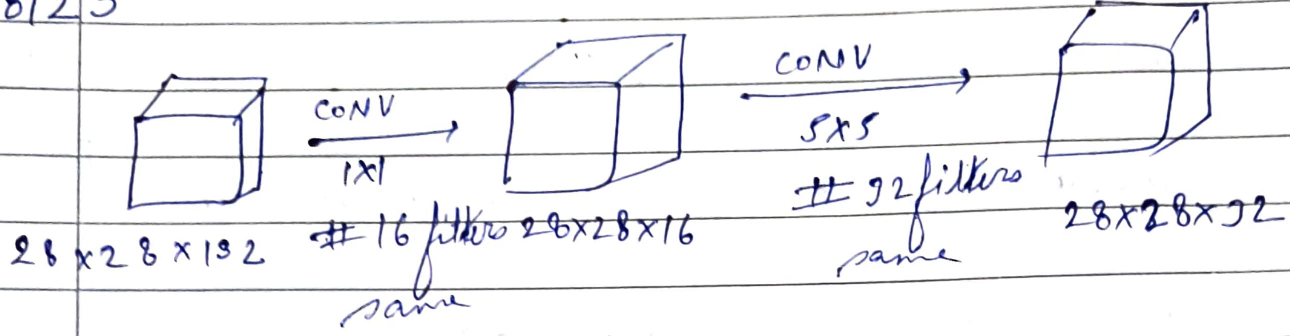


28x28x32

Total operations (mult) = 28x28x192x32 = 3538944

1204 22,400

28/8/25



28x28x192 x 1x1 x 16

28x28x16 x 5X5x32

1) 2408448
2) 10,035,200
3) 12443648 ÷

Modify height width → padding
→ striding, pool

Modify channels → filters
→ 1X1 CONV

Inception Network → some many inception module