Worksheet.

① O/P of Last FC layer.

$$\begin{bmatrix} 2.0 \\ 1.0 \\ 0.1 \end{bmatrix} \xrightarrow[\text{softmax}]{\text{apply}}$$

Softmax formula :- $S(x_i) = \dfrac{e^{x_i}}{\sum_{i=1}^{n} e^{x_i}} = \dfrac{7.39}{7.39 + 2.72 + 1.11}$

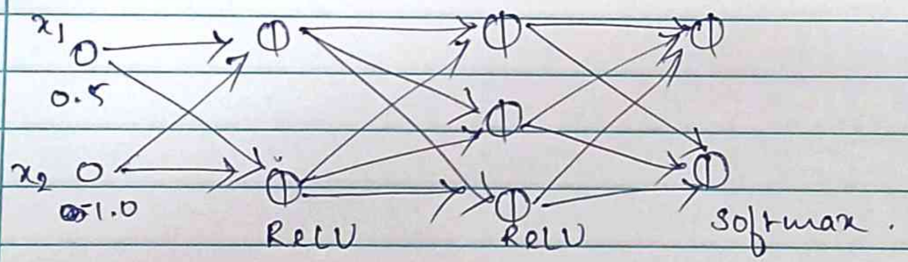
$$= \frac{7.39}{11.22}$$

$P(\hat{y}=0) = 0.65$

$P(\hat{y}=1) = \dfrac{2.72}{11.22} = 0.24$

$P(\hat{y}=2) = \dfrac{1.11}{11.22} = 0.91$

②



$$W^{[1]} = \begin{bmatrix} 0.742 & 0.794 \\ 0.561 & 1.0 \end{bmatrix}$$

$$W^{[2]} = \begin{bmatrix} 0.725 & 0.613 \\ 0.416 & 0.092 \\ 0.876 & 0.590 \end{bmatrix}$$

$$W^{[3]} = \begin{bmatrix} 0.882 & 0.127 & 0.233 \\ 0.112 & 0.132 & 0.536 \end{bmatrix}$$

$$z^{[1]} = W^{[1]}x = \begin{bmatrix} 0.724 & 0.794 \\ 0.561 & 1.0 \end{bmatrix} \begin{bmatrix} 0.5 \\ -1.0 \end{bmatrix} = \begin{bmatrix} -0.423 \\ -0.7195 \end{bmatrix}$$

$a^{[1]} = ReLU(z^{[1]}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$z^{[2]} = w^{[2]} a^{[1]}$

$z^{[2]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$a^{[2]} = ReLU(z^{[2]}) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$z^{[3]} = w^{[3]} \cdot a^{[3]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$a^{[3]} = softmax(z^{[3]})$

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

3)

Forward Pass.

x_1	x_2	$z_1^{[1]}$	$a_1^{[1]}$
0	0	0	0
0	1	1	1
1	0	1	1
1	1	2	1

x_1	x_2	$z_2^{[1]}$	$a_2^{[1]}$
0	0	-1	0
0	1	0	0
1	0	0	0
1	1	1	1

Output layer.

$a_1^{[1]}$	$a_2^{[1]}$	$z^{[2]}$	$a^{[2]}$
0	0	0	0
1	0	1	1
1	0	1	1
1	1	-1	0

$\downarrow$ loss.

$$LOSS(MSE) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

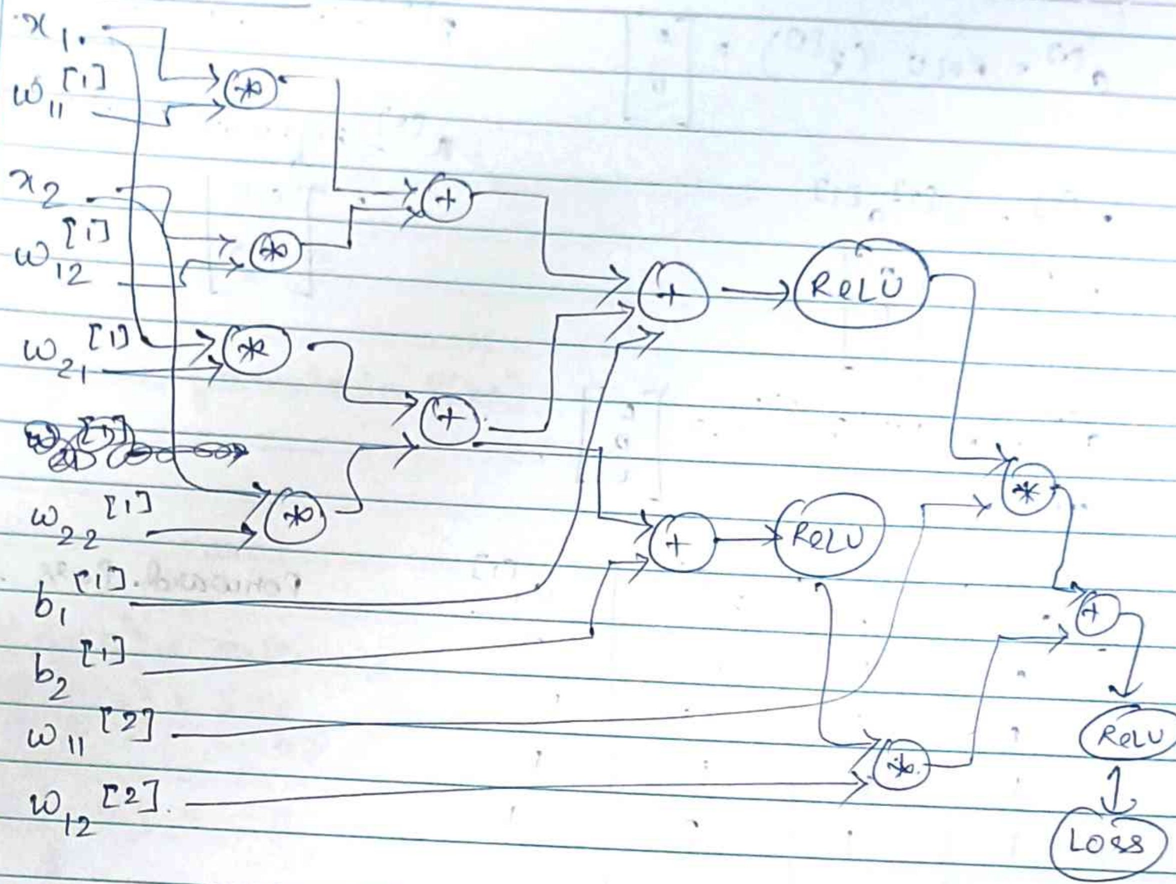
$\hat{y}$	y	L
0	0	0
1	1	0
1	1	0
0	0	0

Total loss = 0.

apsara

Date: __/__/____

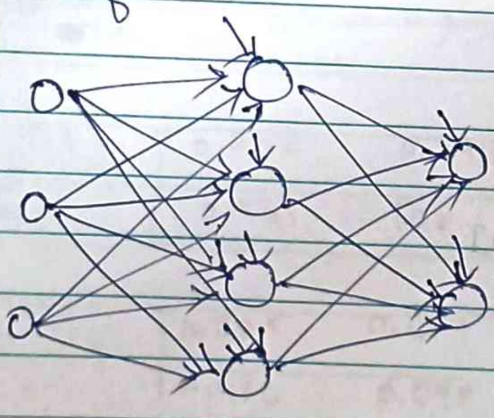
Backward Pass.



$$\frac{dL}{\delta y} = 2(y - \hat{y}) = 0.$$

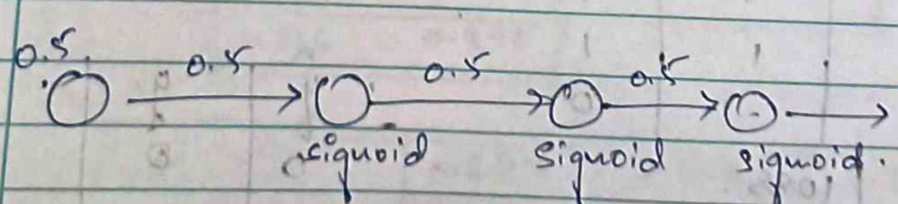
$$\frac{\delta L}{d y} = 0.$$

4)



Total parameters =
$3 \times 4 + 4 + 4 \times 2 + 2$
$= 12 + 4 + 8 + 2$
$= 26$ → parameters in network.

5)



$z^{[1]} = 0.25$	$z^{[2]} = 0.562 \times 0.5 = 0.28$	$z^{[3]} = 0.285$
$a^{[1]} = 0.562$	$a^{[2]} = 0.569$	$a^{[3]} = 0.570$

apsara

Date:

$z^{[1]} = a^{[0]} w^{[1]}$ $a^{[1]} = \text{sigmoid}(z^{[1]})$.

$z^{[2]} = a^{[1]} w^{[2]}$ $a^{[2]} = \text{sigmoid}(z^{[2]})$.

$z^{[3]} = a^{[2]} w^{[3]}$. $a^{[3]} = \text{sigmoid}(z^{[3]})$.

i) $\frac{dy}{dy} = 1$.

ii) $\frac{\delta y}{\delta z^{[3]}} = \frac{\delta y}{\delta y} \cdot \frac{\delta y}{dz^{[3]}} = 1 \times z^{[3]}(1 - z^{[3]}) = 0.285(1 - 0.285)$
$= 0.204$.

iii) $\frac{\delta y}{\delta a^{[2]}} = \frac{\delta y}{\delta z^{[3]}} \cdot \frac{\delta z^{[3]}}{\delta a^{[2]}} = 0.204 \times w^{[3]} = 0.204 \times 0.5 = 0.1018$.

iv) $\frac{\delta y}{\delta z^{[2]}} = \frac{\delta y}{\delta a^{[2]}} \cdot \frac{\delta a^{[2]}}{\delta z^{[2]}} = 0.1018 \times z^{[2]}(1 - z^{[2]}) = 0.1018 \times 0.25 \times 0.72$
$= 0.0205$.

v) $\frac{\delta y}{\delta a^{[1]}} = \frac{\delta y}{\delta z^{[2]}} \cdot \frac{\delta z^{[2]}}{\delta a^{[1]}} = 0.0205 \times 0.5 = 0.01020$.

vi) $\frac{\delta y}{\delta z^{[1]}} = 0.01026 \times 0.28 \times 0.75 = 0.0019240$.

$\frac{\delta y}{\delta w^{[1]}} = 0.0992 \times 0.5 = 0.00096$.

6) $x = [2, 4, 6, 8, 10]$.

$\gamma = 2$.

$\beta = 1$.

x	$(x - \bar{x})^2$	$\frac{x - \bar{x}}{\sqrt{var(x)}}$	$x_{new} = \gamma x_{norm} + \beta$
2	16	-1.265	-1.53
4	4	-0.633	-0.26
6	0	0	1
8	4	0.633	2.66
10	16	1.265	3.53
$\Sigma x = 30$	40		

$$\frac{\Sigma x}{m} = \frac{30}{5} = 6 = \bar{x}$$

$$var(x) = \frac{1}{m-1} \Sigma (x-\bar{x})^2 = \frac{40}{4} = 10.$$

$var(x) = 10.$

7) $0 \xrightarrow{0.5}$ ◯ $\rightarrow$
ReLU
Drop-prob = 0.5

	I/P	O/P
Sample 1	10	5
Sample 2	dropped	
Sample 3	dropped	
Sample 4	15	7.5

S1 → 5
S2 → 5
S3 → 7.5
S4 → 7.5

Total o/p = 12.5   Total o/p = 25

mismatch.

$\frac{12.5}{0.5} = 25$ ↳ This is how it matches.

8) O/P of the convolution operation.

$6 \times 6 * 3 \times 3 = 4 \times 4.$

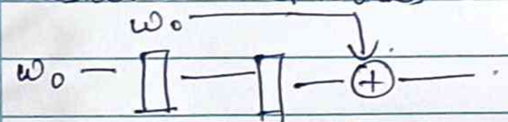
30+0-30 = 0	30+0+0 = 30	30+0+0 = 30	0+0+0 = 0
0	30	30	0
0	30	30	0
0	30	30	0

9) As layers get deeper we encounter vanishing gradient problem - where gradients become very small during backpropagation. This makes it hard for early layers to effectively. Because of vanishing gradients, the

original image features might get diluted or lost over many layers which can hurt training & final performance.

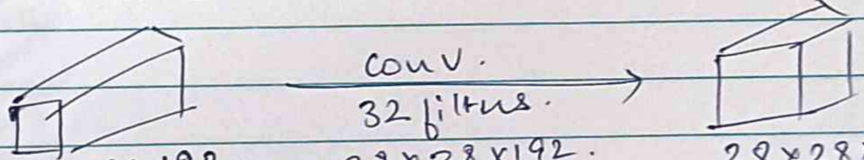
10) The concept of weight sharing automatically reduces the no. parameters to be estimated.
If we take one pixel in the O/P, its obtained only from part of the I/P image & not from the entire image i.e CNNs connect each neuron only to a small local region of the I/P called receptive field unlike FNN where every neuron is connected to one another. This is called Sparsity of connections.

12) Residual networks solve the vanishing gradient problem.



$$\frac{\delta L}{\delta w} = \underset{①}{\frac{\delta L}{\delta w}} + \underset{②}{\frac{\delta L}{\delta w}} \Rightarrow \text{so gradient does not vanish.}$$

13)



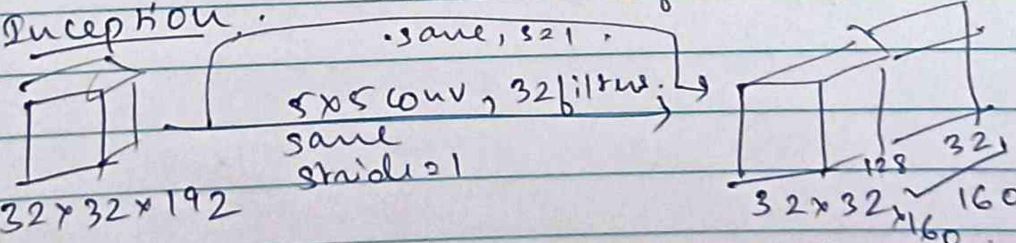
28 × 28 × 192. 28 × 28 × 192. same 28 × 28 × 32.

padding → $\frac{f-1}{2} = p \Rightarrow \frac{28-1}{2} = p$

$p = 13.5$

final dimensions = 55 × 55 × 192.
Total multiplications ⇒ 3,77,64,42,464.

Inception.



1st 3×3 conv = 32 × 32 × 3 × 3 × 128 × 192 = 226492416
2nd 5×5 conv = 32 × 32 × 5 × 5 × 32 × 192 = 157286400
$\overline{383778816}$