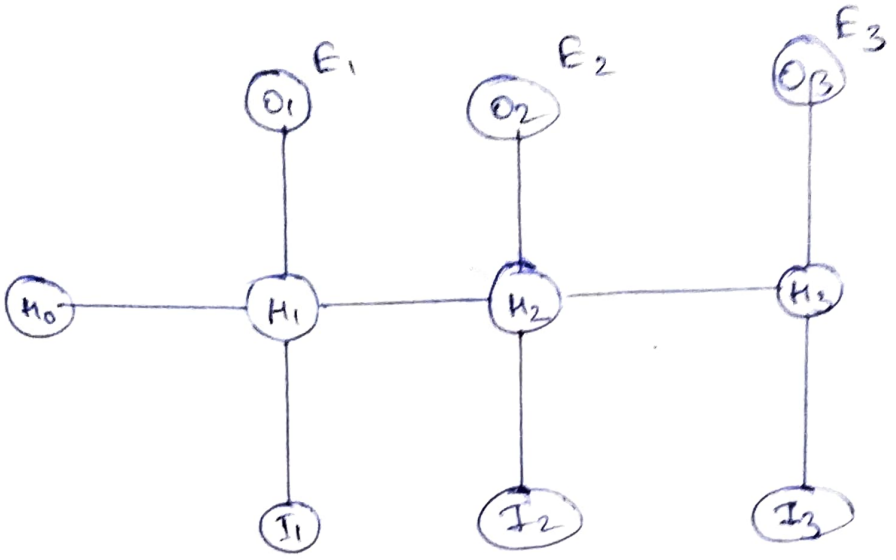
Solution - 1



E = total error

$$E = E_1 + E_2 + E_3 \quad \text{---} \quad (1)$$

(a) (i) $\frac{\partial E}{\partial w} = \frac{\partial E_1}{\partial w} + \frac{\partial E_2}{\partial w} + \frac{\partial E_3}{\partial w}$

(ii) $\frac{\partial E}{\partial U} = \frac{\partial E_1}{\partial U} + \frac{\partial E_2}{\partial U} + \frac{\partial E_3}{\partial U}$

(iii) $\frac{\partial E}{\partial V} = \frac{\partial E_1}{\partial V} + \frac{\partial E_2}{\partial V} + \frac{\partial E_3}{\partial V}$

(b) (i) $\frac{\partial E_3}{\partial w} = \frac{\partial E_2}{\partial O_2} \times \frac{\partial O_2}{\partial H_2} \times \frac{\partial H_2}{\partial w}$

$+$

$$\frac{\partial E_2}{\partial O_2} \times \frac{\partial O_2}{\partial H_2} \times \frac{\partial H_2}{\partial H_1} \times \frac{\partial H_1}{\partial w}$$

(ii) $$\frac{\partial E_2}{\partial U} = \frac{\partial E_2}{\partial O_2} \times \frac{\partial O_2}{\partial H_2} \times \frac{\partial H_2}{\partial U}$$

$$+$$

$$\frac{\partial E_2}{\partial O_2} \times \frac{\partial O_2}{\partial H_2} \times \frac{\partial H_2}{\partial H_1} \times \frac{\partial H_1}{\partial U}$$

(iii) $$\frac{\partial E_2}{\partial V} = \frac{\partial E_2}{\partial O_2} \times \frac{\partial O_2}{\partial H_2} \times \frac{\partial H_2}{\partial V}$$

$$+$$

$$\frac{\partial E_2}{\partial O_2} \times \frac{\partial O_2}{\partial H_2} \times \frac{\partial H_2}{\partial H_1} \times \frac{\partial H_1}{\partial V}$$

(c) (i) $$\frac{\partial E_3}{\partial W} = \frac{\partial E_3}{\partial O_3} \times \frac{\partial O_3}{\partial H_3} \times \frac{\partial H_3}{\partial W}$$

$$+$$

$$\frac{\partial E_3}{\partial O_3} \times \frac{\partial O_3}{\partial H_3} \times \frac{\partial H_3}{\partial H_2} \times \frac{\partial H_2}{\partial W}$$

$$+$$

$$\frac{\partial E_3}{\partial O_3} \times \frac{\partial O_3}{\partial H_3} \times \frac{\partial H_3}{\partial H_2} \times \frac{\partial H_2}{\partial H_1} \times \frac{\partial H_1}{\partial W}$$

(ii) $$\frac{\partial E_3}{\partial U} = \frac{\partial E_3}{\partial O_3} \times \frac{\partial O_3}{\partial H_3} \times \frac{\partial H_3}{\partial U}$$

$$+$$

$$\frac{\partial E_3}{\partial O_3} \times \frac{\partial O_3}{\partial H_3} \times \frac{\partial H_3}{\partial H_2} \times \frac{\partial H_2}{\partial U}$$

$$+$$

$$\frac{\partial E_3}{\partial O_3} \times \frac{\partial O_3}{\partial H_3} \times \frac{\partial H_3}{\partial H_2} \times \frac{\partial H_2}{\partial H_1} \times \frac{\partial H_1}{\partial U}$$

(iii) $$\frac{\partial E_3}{\partial V} = \frac{\partial E_3}{\partial O_3} \times \frac{\partial O_3}{\partial H_3} \times \frac{\partial H_3}{\partial V}$$

$$+$$

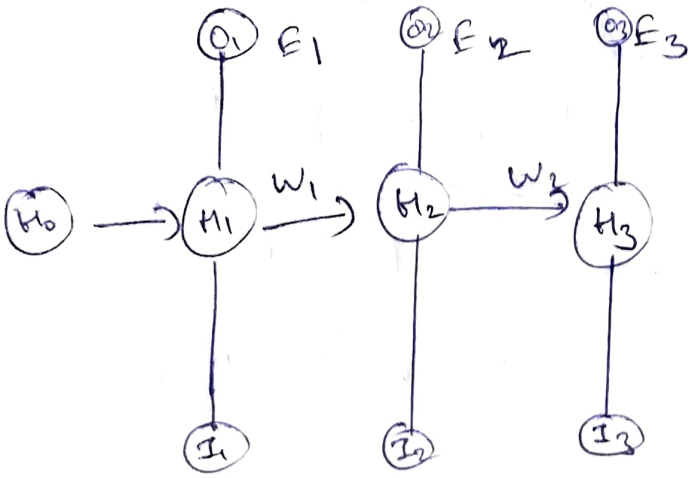
$$\frac{\partial E_3}{\partial O_3} * \frac{\partial O_3}{\partial H_3} * \frac{\partial H_3}{\partial H_2} \times \frac{\partial H_2}{\partial V}$$

$$+$$

$$\frac{\partial E_3}{\partial O_3} \times \frac{\partial O_3}{\partial H_3} \times \frac{\partial H_3}{\partial H_2} \times \frac{\partial H_2}{\partial H_1} \times \frac{\partial H_1}{\partial V}$$

Solution - 2

(a) Why do recurrent models suffer Vanishing Gradient?



In RNN the information travells previous time steps is used for current prediction also, for minimising cost function error is minimised for all the prediction (i.e E_1 E_2 ...)

So the weights matrix needs to be updated. but as the gradient values become so small for initial time steps that the gradient is almost zero and weights are not updated after some epochs Hence this effects the learning

(b) (i) Yes this will be the case of vanishing gradients as the similar frequent words will have similar word embeddings.

Some thing like.

the the the the Use Use Use black Colour ball.

0.2	0.2	0.2	0.2	0.4	0.4	0.4	0.57	0.63	0.8

So at every time step same embedding is passed which makes the gradients to change very minimal. this leads to problem of Vanishing gradient.

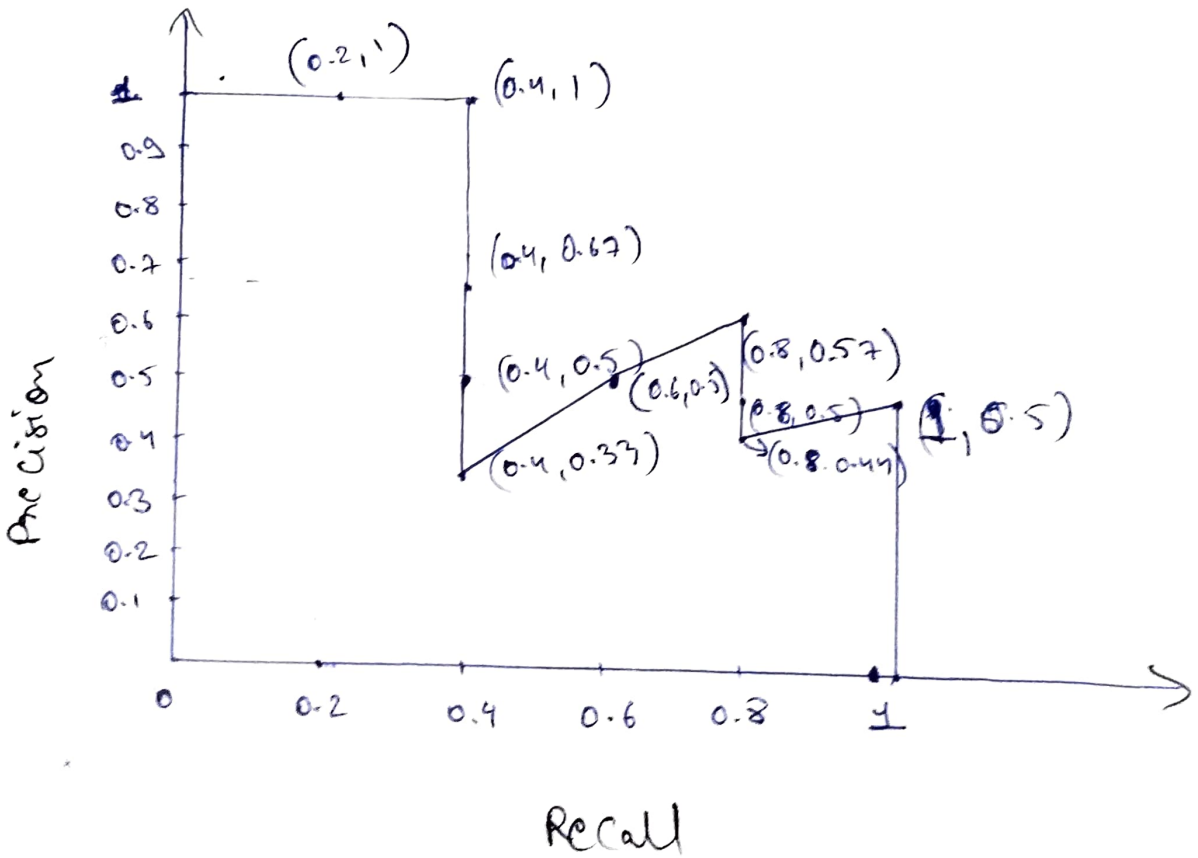
* Also this leads to longer sentences and it becomes hard for our RNN to ~~Here we can use the~~ Store the information from past.

(ii) Solution :- here we can use LSTM which can help store information of past

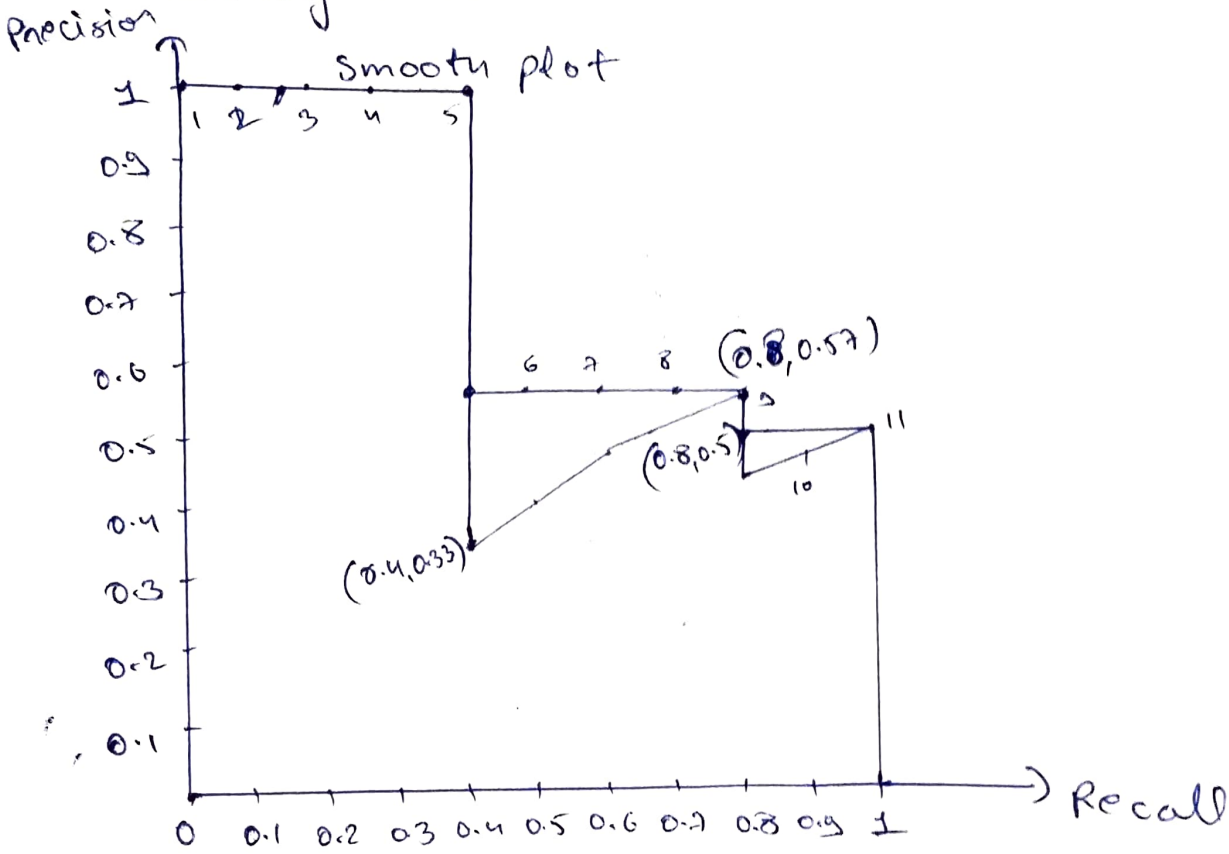
Solution-3

Rank	1	2	3	4	5	6	7	8	9	10
Prediction == correct	True	True	False	False	False	True	True	False	False	True
Precision	$\frac{1}{1}$	$\frac{2}{2}$	$\frac{2}{3}$	$\frac{2}{4}$	$\frac{2}{5}$	$\frac{3}{6}$	$\frac{4}{7}$	$\frac{4}{8}$	$\frac{4}{9}$	$\frac{5}{10}$
Recall	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{5}{5}$

Plotting Precision recall curve.



Smoothing Precision recall curve



Interpolated AP.

AP = mean of values at all 11 points of smooth curve

$$= \frac{(1\times1)+(1\times1)+(1\times1)+(1\times1)+(1\times1)+(1\times0.57)}{11}$$

$$AP = \frac{1+1+1+1+1+0.57+0.57+0.57+0.57+0.5+0.5}{11}$$

$$AP = 0.7527$$

Solution - 4

Cross entropy loss is given as :-

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise} \end{cases}$$

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases}$$

So $CE(p, y)$ can be written as :-

$$CE(p, y) = -\log(p_t). \quad - (1)$$

Focal length is given as :-

$$FL(p_t) = -(1-p_t)^{\gamma} \log(p_t) \quad - (2)$$

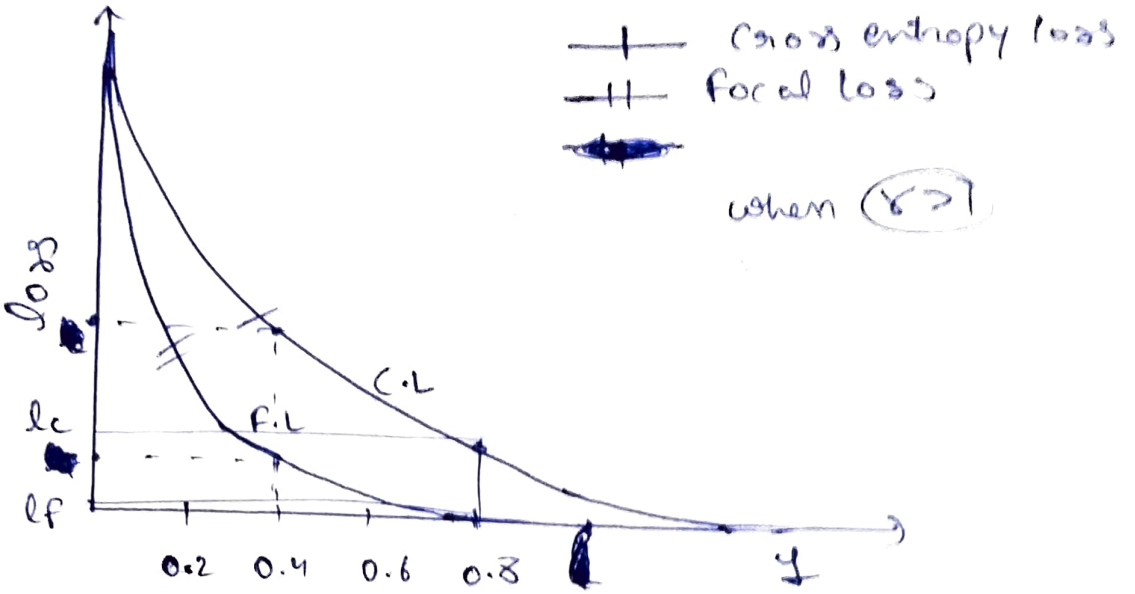
now if $\gamma = 0$.

$$FL(p_t) = -\log p_t \quad - \quad (3)$$

Hence eq. (3) is same as eq^n - (1)

therefore for $\gamma = 0$ in focal loss is equivalent to Cross entropy loss

$$\boxed{(F.L)_{\gamma=0} = C.L}$$

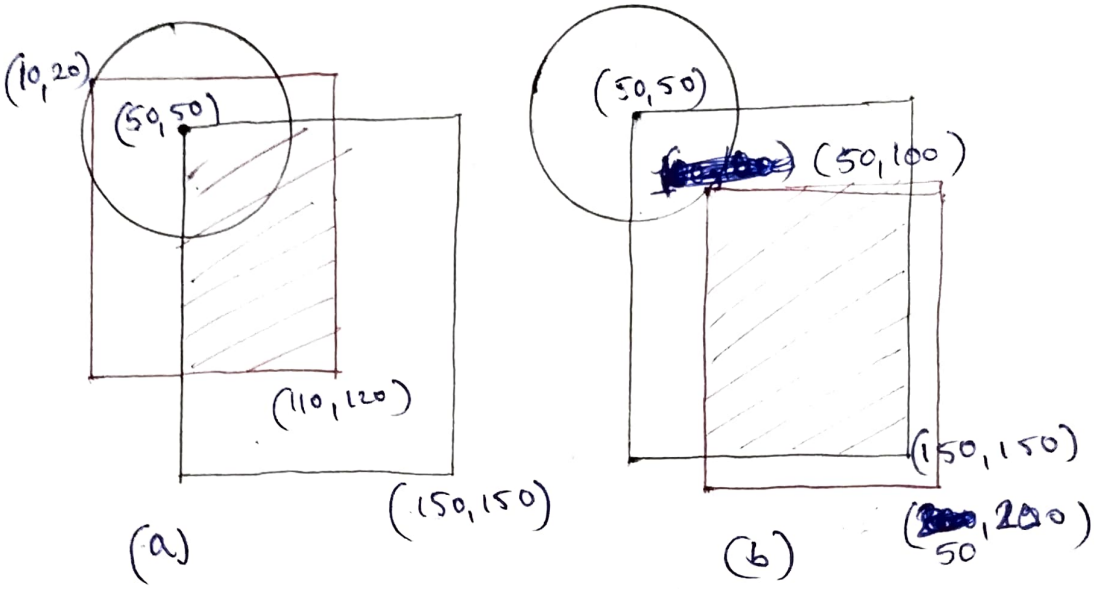


- What focal loss does is for $\gamma > 1$ it reduces the loss for well classified samples (when Probability of prediction > 0.5).

  also for difficult to classify samples (Probability < 0.5) the loss is still very high similar to cross entropy loss.

- Researches show that focal Loss works very well when there is class imbalance in dataset. meaning 1 class is more frequent than other classes.

Solution 5.



To prove that for same L2 norm we can have different IOU values between Ground truth box and Predicted box.

(Case a) → Let us Consider Case (a)

we have two boxes as (50, 50, 150, 150) G.T.B

(10, 20, 110, 120) P.B

- $IOU = \frac{(110-50) \times (120-50)}{(100 \times 100) + (100 \times 100) - ((110-50) \times (120-50))}$
- $\boxed{IOU = 0.2658}$
- $L_2 \text{ Norm} = \sqrt{(50-10)^2 + (50-20)^2} = 50$

Case (b) For case (b)

we have two boxes as

Ground truth box (GTB) = $(50, 50, 150, 150)$

Predicted box (PB) = $(50, 100, 150, 200)$

intersection box = $(50, 100, 150, 150)$

$$\text{IOU} = \frac{(150-50) \times (150-100)}{(100 \times 100) + (100 \times 100) - ((150-50) \times (150-100))}$$

$$\boxed{\text{IOU} = 0.33}$$

$$L_2 \text{ norm} = \sqrt{(50-50)^2 + (100-50)^2}$$

$$L_2 \text{ norm} = 50$$

Hence, we see that for both case (a) & case (b) L_2 norm is equal to 50 but IOU in case (a) is 0.2658 which is different from case (b) i.e 0.33

Intution here is if we fix the shape of both boxes and rotate 1 box with respect to one of the edge of 2nd box. the L_2 norm will remain same but IOU may Change

Solution 6.a

Input (3×3)

1	1	1
1	1	1
1	1	1

3×3

7×7

filter (7×7)

1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

For transposed convolution :-

Output size is given as.

$$\text{output size} = (\text{input size} - 1) \times \text{stride} - 2 \times \text{padding} + (\text{kernel size} - 1) + 1$$

$$\text{Output size} = (3-1) \times 1 - (2 \times 0) + (7-1) + 1$$

$$\text{output size} = 2 + 6 + 1$$

$$\boxed{\text{Output size} = 9}$$

So, output will have a size of 9×9

Output will be = ? with size (9×9)

1×1 = 1	(1×1)+ (1×1) = 2	(1×1)+ (1×1)+ (1×1) = 3	3	3	3	3	2	1
1+1 = 2	2+2 = 4	3+3 = 6	6	6	6	6	4	2
2+1 = 3	4+2 = 6	6+3 = 9	9	9	9	9	6	3
3	6	9	9	9	9	9	6	3
3	6	9	9	9	9	9	6	3
3	6	9	9	9	9	9	6	3
3	6	9	9	9	9	9	6	3
2	4	6	6	6	6	6	4	2
1	2	3	3	3	3	3	2	1

6.b 2D transposed convolution in matrix form

Let us assume

⊙ Kernel =

1	2
3	4

⊙ Input =

1	1
1	1

Output size = (input size − 1) × stride − 2 × padding + (kernel size − 1) + 1

= (2 − 1) × 1 − (2 × 0) + (2 − 1) + 1 = 3

output size = 3 × 3

kernel can be written in form of matrix (W) ↑ width

⊙ no of rows = no of elements in input
(2×2) = 4

⊙ no of columns = no of elements in output
(3×3) = 9

$$W = \begin{bmatrix} 1 & 2 & 0 & 3 & 4 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 3 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 3 & 4 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 3 & 4 \end{bmatrix}_{4\times 9}$$

input can be written as vector form

$$I = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}_{4\times 1}$$

output to transposed convolution will be

$$\boxed{\text{Output} = W^T I}_{9\times 1}$$

$$\text{Output} = \begin{bmatrix} 1 & 3 & 2 & 4 & 10 & 6 & 3 & 7 & 4 \end{bmatrix}$$

⊙ which can be reshaped as :-

$$\text{Output} = \begin{bmatrix} 1 & 3 & 2 \\ 4 & 10 & 6 \\ 3 & 7 & 4 \end{bmatrix}$$