Puneet - Mangla
CS17BTECH11029.

Q.1 a) $$\frac{\partial E}{\partial W} = \frac{\partial E_1}{\partial W} + \frac{\partial E_2}{\partial W} + \frac{\partial E_3}{\partial W}$$

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

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

b) $$\frac{\partial E_2}{\partial W} = \sum_{k=0}^{2} \frac{\partial E_2}{\partial O_2} \frac{\partial O_2}{\partial H_2} \left(\prod_{j=k+1}^{2} \frac{\partial H_j}{\partial H_{j-1}} \right) \frac{\partial H_k}{\partial W}$$

$$\frac{\partial E_2}{\partial U} = \sum_{k=0}^{2} \frac{\partial E_2}{\partial O_2} \cdot \frac{\partial O_2}{\partial H_2} \cdot \left(\prod_{j=k+1}^{2} \frac{\partial H_j}{\partial H_{j-1}} \right) \frac{\partial H_k}{\partial U}$$

$$\frac{\partial E_2}{\partial V} = \frac{\partial E_2}{\partial O_2} \cdot \frac{\partial O_2}{\partial V} = \frac{\partial E_2}{\partial O_2} \cdot \frac{\partial O_2}{\partial Z_2} \cdot \frac{\partial Z_2}{\partial V} \quad \{ Z_2 = VH_2 \}$$

c) $$\frac{\partial E_3}{\partial W} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial O_3} \frac{\partial O_3}{\partial H_3} \left(\prod_{j=k+1}^{3} \frac{\partial H_j}{\partial H_{j-1}} \right) \frac{\partial H_k}{\partial W}$$

$$\frac{\partial E_3}{\partial U} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial O_3} \frac{\partial O_3}{\partial H_3} \left(\prod_{j=k+1}^{3} \frac{\partial H_j}{\partial H_{j-1}} \right) \frac{\partial H_k}{\partial W}$$

$$\frac{\partial E_3}{\partial V} = \frac{\partial E_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial Z_3} \cdot \frac{\partial Z_3}{\partial V} \quad \{ Z_3 = VH_3 \}$$

Q.3

Rank	0	1	2	3	4	5	6	7	8	9	10
precision	1	1	0.6	0.5	0.4	0.5	0.57	0.5	0.44	0.5	
Recall	0.2	0.4	0.4	0.4	0.4	0.6	0.8	0.8	0.8	1	

r	0.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$P_{interp}(r)$	1	1	1	1	1	0.57	0.57	0.57	0.57	0.5	0.5

$$AP_{interp} = \frac{1}{11} \sum_{r} P_{interp}(r)$$

$$= \frac{1}{11}(8.28) = 0.7527$$

Q.2 a) $$\frac{\partial E_T}{\partial W} = \sum_{k=0}^{T} \frac{\partial E_T}{\partial y_T} \cdot \frac{\partial y_T}{\partial h_T} \left(\prod_{j=k+1}^{T} \frac{\partial h_j}{\partial h_{j-1}} \right) \frac{\partial h_k}{\partial W}$$

if $\left| \frac{\partial h_j}{\partial h_{j-1}} \right| < 1$ { might be because of use of sigmoid activations}

then $\left| \prod_{j=k+1}^{T} \frac{\partial h_j}{\partial h_{j-1}} \right|$ will vanish, resulting in

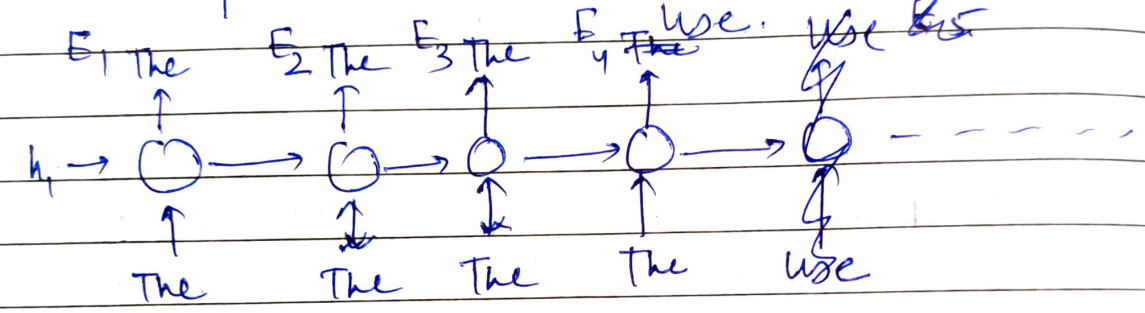
no or very minor gradient of E_T w.r.t W

i.e E_T will not have any contribution in updating weight W.

Solution:- Use of gated networks or activations that are not bounded between -1, 1

b) (i) The data above can be confusing for a normal recurrent network as, ~~many~~ same words are repeated many times before a new word comes.

A model can get confused if a 'the' should follow 'the' or 'use' should follow a 'the'. The model can become biased to former. The reason of this is because of vanishing gradient problem.



E_i represents error.

now while calculating gradient of total Error w.r.t to W, we get following

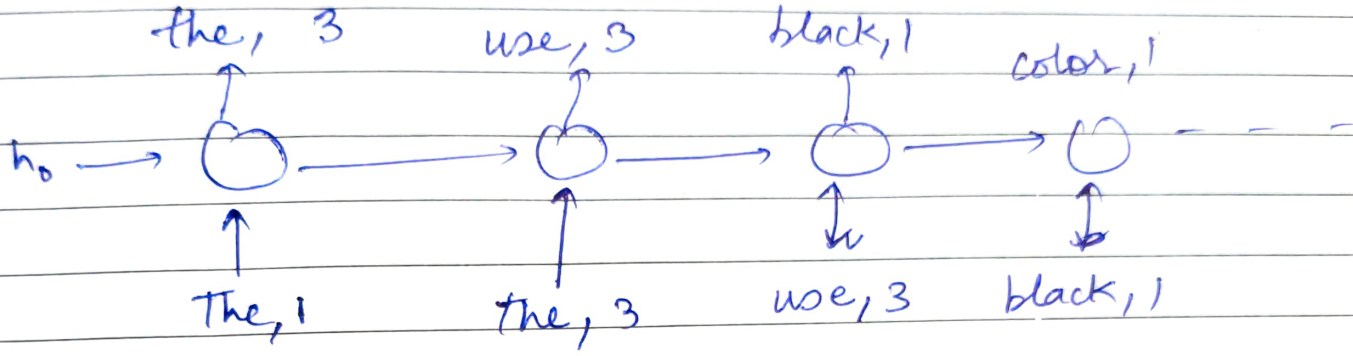
$$E = E_1 + E_2 + E_3 + E_4 \quad \cdots$$

$$\frac{\partial E}{\partial W} = \frac{\partial E_1}{\partial W} + \frac{\partial E_2}{\partial W} + \frac{\partial E_3}{\partial W} + \frac{\partial E_4}{\partial W} \quad \cdots$$

Errors E_1, E_2, E_3 focus on predicting 'the' after seeing 'the', whereas E_4 will focus on predicting 'use' after seeing 'the'

Hence the contribution of E_4 in updating W will be relatively less than E_1, E_2, E_3 which is equivalent to vanishing of $\frac{\partial E_4}{\partial W}$ gradient.

(ii) The suggested modification is :
Along with predicting the final next occurence, predict the no. of times it will appear consecutively before a new word.



Now, if we use gated networks, the vanishing gradient problem will be resolved.

Q.4 $FL = -(1-p)^{\gamma} \log p$

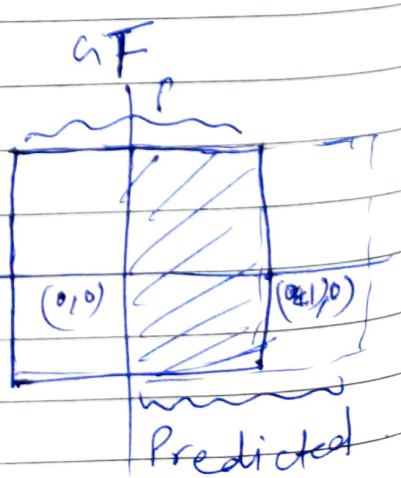
when $\gamma = 0$ $FL = -\log p$ which is same as standard cross Entropy loss.

For correctly classified examples, Focal loss is relatively less than cross entropy loss thus preventing the model to make over-confident predictions which can negatively affect its overall performance.

For incorrectly classified examples, the focal loss is not affected much - just decreased slightly as $(1-p)^{\gamma}$ is more near 1 and in misclassification $p < 0.5$

Q.5 A bounding box can be Represented as 4 dimensional vector - (x, y, h, w)
$x, y \rightarrow$ coordinates of center
$h, w \rightarrow$ height and width of bounding box.
Consider following 2 cases

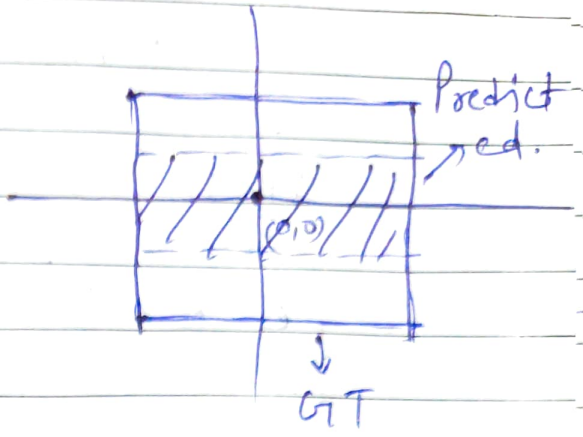
(1) GT bbox = $(0, 0, 2, 2)$
Predicted bbox = $(0, 0, 2, 2)$



L_2 norm is $= \sqrt{1^2 + 0^2 + 0^2 + 0^2}$
$= 1$

$IOU = \frac{2}{6} = \frac{1}{3}$

(2) GT bbox = (0, 0, 2, 2)
Predicted bbox = (0, 0, 1, 2)



L_2 norm = $\sqrt{0^2+0^2+1^2+0^2}$
= 1

IOU = $\frac{2}{4} = \frac{1}{2}$

The reason why it happens is because the L_2 norm of vector representation of a bounding box - $\|(x, y, h, w)\|_2$ doesn't take into the relative shift/position of bboxes. It considers each of the elements - {x, y, h, w} as seperate quantity while calculating the norm.
IOU on other hand, takes into accent the relative position of two bboxes.

Q.6 a) $H_O = (H_I - 1) * \text{stride} + H_k - 2 * \text{padding}$
$W_O = (W_I - 1) * \text{stride} + W_k - 2 * \text{padding}$.
H_O, W_O - height, width of output
H_I, W_I - " " of input
H_k, W_k - " " of kernel

by applying above equations
$H_O = 9$
$W_O = 9$
So size of result is 9x9.

b) $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ $\begin{bmatrix} e & f \\ g & h \end{bmatrix}$

Input ↓ output →

$$\begin{bmatrix} e & 0 & 0 & 0 \\ f & e & 0 & 0 \\ 0 & f & 0 & 0 \\ g & 0 & e & 0 \\ h & g & f & e \\ 0 & h & 0 & f \\ 0 & 0 & g & 0 \\ 0 & 0 & h & 0 \\ 0 & 0 & 0 & h \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} ae \\ af + be \\ bf \\ ag + ec \\ ah + gb + fc + de \\ bh + fd \\ gc \\ hc + dg \\ hd \end{bmatrix}$$

↓

resize into 3x3.

output.