Name - Shivashish Suman

Roll No. - CS17BTECH11037

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

$$\frac{\partial E_3}{\partial W} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial H_3} \cdot \frac{\partial H_3}{\partial H_k} \cdot \frac{\partial H_k}{\partial W}$$

$$\frac{\partial E_2}{\partial W} = \sum_{k=0}^{2} \frac{\partial E_2}{\partial O_2} \cdot \frac{\partial O_2}{\partial H_2} \cdot \frac{\partial H_2}{\partial H_k} \cdot \frac{\partial H_k}{\partial W}$$

$$\frac{\partial E_1}{\partial W} = \sum_{k=0}^{1} \frac{\partial E_1}{\partial O_1} \cdot \frac{\partial O_1}{\partial H_1} \cdot \frac{\partial H_1}{\partial H_k} \cdot \frac{\partial H_k}{\partial W}$$

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

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

$$= \frac{\partial E_3}{\partial O_3} \cdot H_3 = (O_3 - \hat{O}_3) \cdot H_3$$

{Considering mean squared error. $\hat{O}_3$ is ground truth value.

Similarly,

$$\frac{\partial E_2}{\partial V} = (O_2 - \hat{O}_2) \cdot H_2$$

$$\frac{\partial E_1}{\partial V} = (O_1 - \hat{O}_1) \cdot H_1$$

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

$$\frac{\partial E_3}{\partial U} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial H_3} \cdot \frac{\partial H_3}{\partial H_k} \cdot \frac{\partial H_k}{\partial U}$$

$$\frac{\partial E_2}{\partial U} = \sum_{K=0}^{2} \frac{\partial E_2}{\partial O_2} \cdot \frac{\partial O_2}{\partial H_2} \cdot \frac{\partial H_2}{\partial H_K} \cdot \frac{\partial H_K}{\partial U}$$

$$\frac{\partial E_1}{\partial U} = \sum_{K=0}^{1} \frac{\partial E_1}{\partial O_1} \cdot \frac{\partial O_1}{\partial H_1} \cdot \frac{\partial H_1}{\partial H_K} \cdot \frac{\partial H_K}{\partial U}$$

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

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

$$\frac{\partial E_2}{\partial V} = (O_2 - \hat{O}_2) \cdot H_2$$

(c). $$\frac{\partial E_3}{\partial W} = \sum_{K=0}^{2} \frac{\partial E_{3}}{\partial O_3} \cdot \frac{\partial O_3}{\partial H_3} \left(\prod_{j=K+1}^{3} \frac{\partial H_j}{\partial H_{j-1}} \right) \cdot \frac{\partial H_K}{\partial W}$$

$$\frac{\partial E_3}{\partial U} = \sum_{K=0}^{2} \frac{\partial E_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial H_3} \left(\prod_{j=K+1}^{3} \frac{\partial H_j}{\partial H_{j-1}} \right) \frac{\partial H_K}{\partial U}$$

$$\frac{\partial E_3}{\partial V} = (O_3 - \hat{O}_3) \cdot H_3$$

2. (a) Recurrent models suffer from vanishing gradient because for sigmoid activations, gradient is upper bounded by 1. So if we have product of gradients, $g_1, g_2 \cdot g_n$ if all are < 1 then product will decrease which will lead to vanishing gradient.

We can solve this problem by using gradient highway as we have it in LSTM which uses cell state and hidden state.

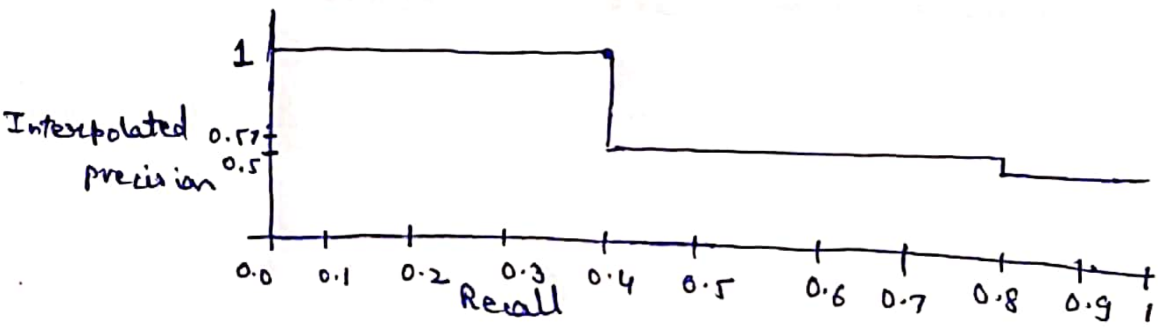
(b). (i) In this dataset, there are multiple repeated and irrelevant words which can lead to increase in the sequence length and hence will lead to vanishing gradient as gradients have to be propogated back through larger length. Hence, the problem can be related to the vanishing gradient problem.

(ii) LSTM can help to solve the problem as there is a gradient highway which passes the gradient without diminishing to deeper layers and hence LSTM will be helpful which will give less weight to the repeated words.

3. We know that $\text{Precision} = \frac{TP}{TP+FP}$ and $\text{recall} = \frac{TP}{TP+FN}$

So we calculate precision and recall.

Rank	Is a correct prediction?	Precision	Recall
1	True	$1/1 = 1.0$	$1/5 = 0.2$
2	True	$2/2 = 1.0$	$2/5 = 0.4$
3	False	$2/3 = 0.67$	$2/5 = 0.4$
4	False	$2/4 = 0.5$	$2/5 = 0.4$
5	False	$2/5 = 0.4$	$2/5 = 0.4$
6	True	$3/6 = 0.5$	$3/5 = 0.6$
7	True	$4/7 = 0.57$	$4/5 = 0.8$
8	False	$4/8 = 0.5$	$4/5 = 0.8$
9	False	$4/9 = 0.44$	$4/5 = 0.8$
10	True	$5/10 = 0.5$	$5/5 = 1.0$



(Interpolated Precision at r) is $\max_{\tilde{r} \geq r} (\text{Precision at } \tilde{r})$.

$$\therefore \text{Interpolated AP} = \frac{1}{11}(5 \times 1.0 + 4 \times 0.57 + 2 \times 0.5)$$

$$= 0.752$$

$\therefore$ Interpolated AP is 0.752.

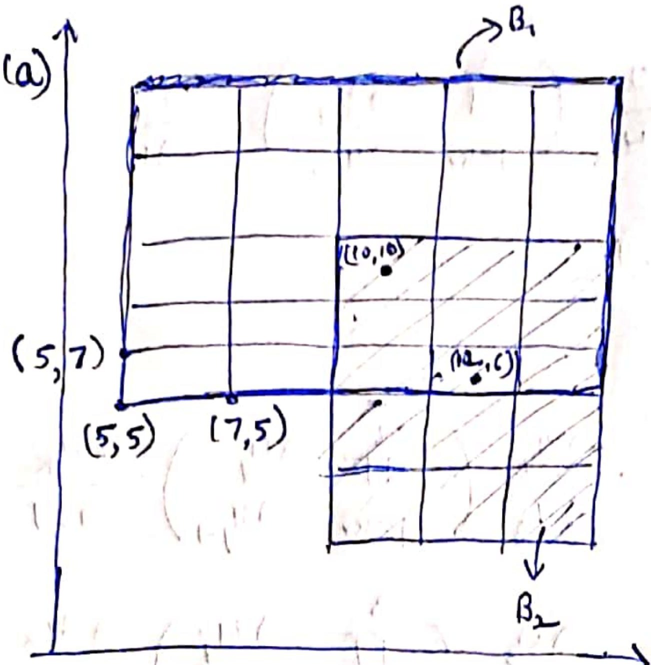
4. At $\gamma = 0$, $FL(p) = -\log p$

and we know that cross entropy loss is $-\log p$.

$\therefore$ If we set $\gamma = 0$, then focal loss is equal to the standard cross entropy loss.

So in this case, ~~St~~ even when $p >> 0.5$ i.e. ~~eas~~ sample is getting classified correctly, we observe that the loss incurred is high. for $\gamma > 0$, if sample is already classified correctly then its contribution to the loss decreases. But it does not happen when $\gamma = 0$.

5. (a)



{Each small box is of size (2x2).}

Let B_2 be the ~~second~~ predicted bounding box and B_1 be the ground truth.

Then L2 norm between B_1 and ~~B_2~~ B_2 will be –

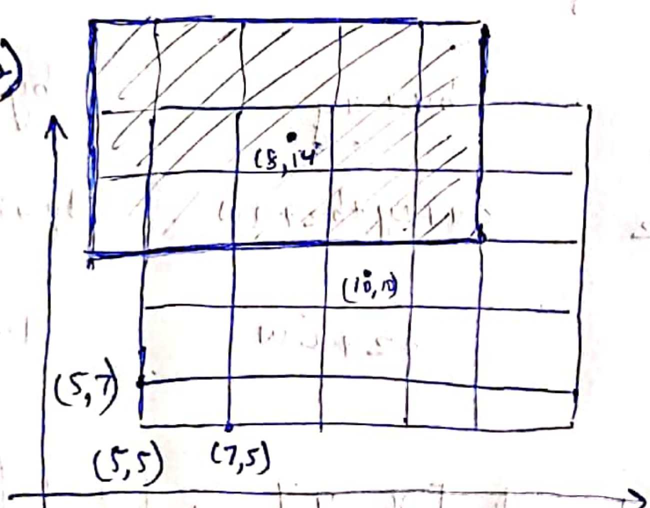
$$\sqrt{(12-10)^2 + (10-6)^2}$$

$$= 2\sqrt{5} = 4.47$$

$$IOU = \frac{4 * 9}{100 + 60 - 36} = \frac{36}{124}$$

$$= \frac{9}{31} = 0.29$$

(b)



centre of B_2 is at (8, 14) and it is of size (10x6)

L2 norm is $\sqrt{(10-8)^2 + (14-10)^2}$

$$= 2\sqrt{5} = 4.47$$

$$IOU = \frac{4 * 8}{100 + 60 - 32} = \frac{32}{128}$$

$$= 0.25$$

which is different from above.

Thus we have shown that we can have same L_2-norm between two bounding boxes with differing IOU values. An explanation is that L_2 norm does not give information about how the boxes are oriented and hence two boxes with same L_2-norm can have different IOUs depending on orientation.

6. (a) Output will be of size $(3+7-1) * (3+7-1)$

$$= 9 * 9$$

(b) Let $\begin{bmatrix} x & y \\ z & w \end{bmatrix}$ be 2×2 input and $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ be 2×2 kernel.

Transposed convolution can be expressed as matrix multiplication.

$$\begin{bmatrix} a & 0 & 0 & 0 \\ b & a & 0 & 0 \\ 0 & b & 0 & 0 \\ c & 0 & a & 0 \\ d & c & b & a \\ 0 & d & 0 & b \\ 0 & 0 & c & 0 \\ 0 & 0 & d & c \\ 0 & 0 & 0 & d \end{bmatrix}_{9 \times 4} * \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix}_{4 \times 1} = \begin{bmatrix} ax \\ bx + ay \\ by \\ cx + az \\ dx + cy + bz + aw \\ dy + bw \\ cz \\ dz + cw \\ dw \end{bmatrix}_{9 \times 1}$$

After getting (9×1) matrix, we convert it into 3×3 matrix by taking first 3 elements in 1^{st} row, next 3 in 2^{nd} row and last 3 in 3^{rd} row.

So final output will be $\begin{bmatrix} ax & bx + ay & by \\ cx + az & dx + cy + bz + aw & dy + bw \\ cz & dz + cw & dw \end{bmatrix}_{3 \times 3}$

which is transposed convolution of $\underbrace{\begin{bmatrix} x & y \\ z & w \end{bmatrix}_{2 \times 2}}_{\text{input}}$ and $\underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}_{2 \times 2}}_{\text{Kernel}}$.