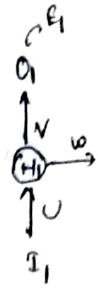
DLV Assignment - 4

Vinta Reethu
ES18BTECH11028

1. We know that

$$h_t = \tanh[UI_t + Wh_{t-1}]$$
$$\hat{y}_t = \text{softmax}[Vh_t]$$



- Let us assume rent loss is used then we can write the derivative of E_t wrt h_t as given below

$$\frac{\partial E_t}{\partial h_t} = \frac{\partial E_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial z_t} \cdot \frac{\partial z_t}{\partial h_t}$$
$$= (\hat{y}_t - y_t)V$$

- Calculating derivative of h_t,

$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial}{\partial h_{t-1}}[Wh_{t-1} + UI_t]$$
$$= W(1 - \tanh^2[UI_t + Wh_{t-1}])$$
$$= W(1 - h_t^2)$$

- Calculating $\frac{\partial E_1}{\partial W}, \frac{\partial E_1}{\partial U}, \frac{\partial E_1}{\partial V}$:

$$\frac{\partial E_1}{\partial W} = \frac{\partial E_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial W}$$
$$= (\hat{y}_1 - y_1) V (1 - h_1^2) h_0$$

$$\frac{\partial E_1}{\partial U} = \frac{\partial E_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial U}$$
$$= (\hat{y}_1 - y_1) V (1 - h_1^2) I_1$$

$$\frac{\partial E_1}{\partial V} = \frac{\partial E_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial V}$$
$$= (\hat{y}_1 - y_1) h_1$$

b) Calculating $\frac{\partial E_2}{\partial W}, \frac{\partial E_2}{\partial U}, \frac{\partial E_2}{\partial V}$:

$$\frac{\partial E_2}{\partial W} = \frac{\partial E_2}{\partial h_2} \sum_{k=1}^{2} \left(\frac{\partial h_2}{\partial h_k} \cdot \frac{\partial h_k}{\partial W} \right)$$

$$= (\hat{y}_2 - y_2) \cdot V \cdot \left(\frac{\partial h_2}{\partial W} + \left[\frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W} \right] \right)$$

$$= (\hat{y}_2 - y_2) \cdot V \cdot \left[((1 - h_2^2) h_1) + [(1 - h_2^2) W (1 - h_1^2) h_0] \right]$$

$$= (\hat{y}_2 - y_2) \cdot V \cdot (1 - h_2^2) \cdot [h_1 + W(1 - h_1^2) h_0]$$

$$\frac{\partial E_2}{\partial U} = \frac{\partial E_2}{\partial h_2} \sum_{k=1}^{2} \left[\frac{\partial h_2}{\partial h_k} \cdot \frac{\partial h_k}{\partial W} \right]$$

$$= (\hat{y}_2 - y_2) \cdot V \cdot \left(\frac{\partial h_2}{\partial U} + \left[\frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial U} \right] \right)$$

$$= (\hat{y}_2 - y_2) \cdot V \cdot \left[(1 - h_2^2) I_2 + (1 - h_2^2) W (1 - h_1^2) I_1 \right]$$

$$= (\hat{y}_2 - y_2) \cdot V \cdot (1 - h_2^2) \left[I_2 + W (1 - h_1^2) I_1 \right]$$

$$\frac{\partial E_2}{\partial V} = \frac{\partial E_2}{\partial Z_2} \cdot \frac{\partial Z_2}{\partial V}$$

$$= (\hat{y}_2 - y_2) \cdot h_2$$

c) Calculating $\frac{\partial E_3}{\partial W}, \frac{\partial E_3}{\partial U}, \frac{\partial E_3}{\partial V}$:

$$\frac{\partial E_3}{\partial W} = \frac{\partial E_3}{\partial h_3} \sum_{k=1}^{3} \left[\frac{\partial h_3}{\partial h_k} \cdot \frac{\partial h_k}{\partial W} \right)$$

$$= (\hat{y}_3 - y_3) \cdot V \cdot \left(\frac{\partial h_3}{\partial W} + \left[\frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial W} \right] + \left[\frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W} \right] \right)$$

$$= (\hat{y}_3 - y_3) \cdot V \cdot \left((1 - h_3^2) h_2 + (1 - h_3^2) W (1 - h_2^2) h_1 + (1 - h_3^2) W (1 - h_2^2) W (1 - h_1^2) h_0 \right]$$

$$= (\hat{y}_3 - y_3) \cdot V \cdot (1 - h_3^2) \left[h_2 + W (1 - h_2^2) \left(h_1 + W (1 - h_1^2) h_0 \right) \right]$$

$$\frac{\partial E_3}{\partial U} = \frac{\partial E_3}{\partial h_3} \sum_{k=1}^{3} \left[\frac{\partial h_3}{\partial h_k} \cdot \frac{\partial h_k}{\partial U} \right]$$

$$= (\hat{y}_3 - y_3) \cdot V \cdot \left(\frac{\partial h_3}{\partial U} + \left[\frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial U} \right] + \left[\frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial U} \right] \right)$$

$$= (\hat{y}_3 - y_3) \cdot V \cdot \left((1 - h_3^2) \underset{b(I_3)}{\cancel{I}_3} + (1 - h_3^2) W (1 - h_2^2) I_2 + (1 - h_3^2) W (1 - h_2^2) W (1 - h_1^2) I_1 \right)$$

$$= (\hat{y}_3 - y_3) \cdot V \cdot (1-h_3^2) \cdot \left(I_3 + W(1-h_2^2)\left(I_2 + W(1-h_1^2) I_1\right)\right)$$

$$\frac{\partial E_3}{\partial V} = \frac{\partial E_3}{\partial z_3} \cdot \frac{\partial z_3}{\partial V}$$

$$= (\hat{y}_3 - y_3) \cdot h_3$$

a)

$$\frac{\partial E}{\partial W} = \sum_{K=1}^{3} \frac{\partial E_K}{\partial W}$$

$$= (\hat{y}_1 - y_1) V (1-h_1^2) h_0 +$$
$$(\hat{y}_2 - y_2) V (1-h_2^2) [h_1 + W(1-h_1^2) h_0] +$$
$$(\hat{y}_3 - y_3) V (1-h_3^2) [h_2 + W(1-h_2^2)(h_1 + W(1-h_1^2) h_0)]$$

$$\frac{\partial E}{\partial U} = \sum_{K=1}^{3} \frac{\partial E_K}{\partial U}$$

$$= (\hat{y}_1 - y_1) V (1-h_1^2) I_1 +$$
$$(\hat{y}_2 - y_2) V (1-h_2^2) [I_2 + W(1-h_1^2) I_1] +$$
$$(\hat{y}_3 - y_3) V (1-h_3^2) [I_3 + W(1-h_2^2)(I_2 + W(1-h_1^2) I_1)]$$

$$\frac{\partial E}{\partial V} = \sum_{K=1}^{3} \frac{\partial E_K}{\partial V}$$

$$= (\hat{y}_1 - y_1) h_1 +$$
$$(\hat{y}_2 - y_2) h_2 +$$
$$(\hat{y}_3 - y_3) h_3$$

a) If we look at the gradient term of E wrt W,

$$\frac{\partial E_3}{\partial W} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial \hat{y}_3} \cdot \frac{\partial \hat{y}_3}{\partial h_3} \left(\prod_{j=k+1}^{3} \frac{\partial h_j}{\partial h_{j-1}} \right) \frac{\partial h_k}{\partial W}$$

For sigmoid activation functions, this gradient is upper bounded by 1. The gradients that carry information becomes smaller and smaller and will vanish over time. Hence long term dependencies will not contribute for learning.

We can solve this vanishing gradient problem in RNN by

- Using ReLU as activation function.
- Using Regularization.
- Better initialization of weights.
- Using short time sequences.
- Using architectures such as LSTMs and GRUs

b) i) The above given dataset has repetitive words. If there are words that are repeated in almost entire dataset, it has to capture the long term dependencies inorder to predict accurately. So to capture the long term dependencies. the network has to back propogate through time over a long interval. As discussed above this would lead to vanishing gradient problem.

ii) We can use ReLU as an activation function or we can use regularization. We could use LSTM's instead of RNN's as it proven to avoid vanishing gradient problem (as it has cell states connected by highway).

3. Let us look at some of the formulaes that are needed to compute interpolated AP

- Precision $= \frac{CTP}{CTP + CFP}$ c - cummulative

- Recall $= \frac{CTP}{CTP + CFN} = \frac{CTP}{\text{Bounding boxes in total}}$

CTP	CFP	Rank	Precision	Recall
0	0	0	0	0
1	0	1	1	$\frac{1}{5}$
2	0	2	1	$\frac{2}{5}$
2	1	3	$\frac{2}{3}$	$\frac{2}{5}$
2	2	4	$\frac{1}{2}$	$\frac{2}{5}$
2	3	5	$\frac{2}{5}$	$\frac{2}{5}$
3	3	6	$\frac{1}{2}$	$\frac{3}{5}$
4	3	7	$\frac{4}{7}$	$\frac{4}{5}$
4	4	8	$\frac{1}{2}$	$\frac{4}{5}$
4	5	9	$\frac{4}{9}$	$\frac{4}{5}$
5	5	10	$\frac{1}{2}$	1

- The formula for calculating interpolated AP is

$$AP = \frac{1}{11} \sum_{r \in (0, 0.1, \ldots 1)} P_{interp}(r)$$

where

$$P_{interp}(r) = \max_{\tilde{r} : \tilde{r} \geq r} p(\tilde{r})$$

- Therefore $P_{interp}(r)$ can be as below

Recal level	$P_{interp}(r)$
0	1
0.1	1
0.2	1
0.3	1
0.4	1
0.5	$\frac{4}{7}$
0.6	$\frac{4}{7}$
0.7	$\frac{4}{7}$
0.8	$\frac{4}{7}$
0.9	$\frac{1}{2}$
1	$\frac{1}{2}$

Interpolated AP $= \frac{1}{11}\left[1+1+1+1+1+1+\frac{4}{7}+\frac{4}{7}+\frac{4}{7}+\frac{4}{7}+\frac{1}{2}+\frac{1}{2}\right]$

$= \frac{1}{!} \frac{58}{77} = \underline{\underline{0.753}}$

4. • Focal loss function is given by $-(1-p)^{\gamma}\log p$.
   • The standard cross entropy loss is given by $-\log p$.
   • When $\gamma = 0$, focal loss is equal to standard cross entropy loss.
   • The way focal loss work is, it increases (or penalizes) loss for incorrectly classified data and decreases loss for correctly classified data.

5. We can prove the claim as follows.
   • Let (x_1, y_1, x_2, y_2) be a tuple denoting a bounding box.
   • Lets fin the distance between one of the two corners say p. Consider a circle of radius r around the other corner of the ground truth bounding box.
   • For any predicted bounding box with corresponding opposite corner lying on this circle will have the same L_2 norm of $\sqrt{p^2+r^2}$. Despite of having same L_2 norm these boxes can have different IOU's.
   • This can happen because L_2 norm is a distance measure, whereas IoU measures the extent of overlap between the bounding boxes.

6. a) Given that, input_size = 3×3
   filter = 7×7
   stride = 1
   padding = 0

We know that the output shape of transpose convolution is

input_size × stride − stride + filter − 2 × padding

Therefore, output_size $= 3 \times 1 - 1 + 7 - 2(0)$

$= 3 - 1 + 7$

$= 9$

∴ output → 9×9 matrix

b) Input = 2x2
Kernel = 2x2

For example let us consider,

$$\text{Input} = \begin{bmatrix} 25 & 31 \\ 43 & 49 \end{bmatrix}$$

$$\text{Kernel} = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 0 & 3 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 25 \\ 31 \\ 43 \\ 49 \end{bmatrix} = \begin{bmatrix} 0 \\ 1\cdot 25 \\ 1\cdot 31 \\ 2\cdot 25 \\ 3\cdot 25 + 2\cdot 31 + 1\cdot 43 \\ 3\cdot 31 + 1\cdot 49 \\ 2\cdot 43 \\ 3\cdot 43 + 2\cdot 49 \\ 3\cdot 49 \end{bmatrix}$$

Transposed weight matrix

$$= \begin{bmatrix} 0 \\ 25 \\ 31 \\ 50 \\ 180 \\ 142 \\ 86 \\ 227 \\ 147 \end{bmatrix} \overset{\text{reshape}}{=} \begin{bmatrix} 0 & 25 & 31 \\ 50 & 180 & 142 \\ 86 & 227 & 147 \end{bmatrix}$$

- Inorder to achieve the transposed weight matrix, we can use the following code snippet.

```
def matrix_transposed(Filter):
    K,W = torch.zeros(5), torch.zeros((4,9))
    K[:2], K[3:5] = filter[0,:], K[1,:]
    W[0,:5], W[1,1:6], W[2,3:8], W[3,4:] = K, K, K, K
    return W.T
```

- Once you get the above transposed matrix we just have to multiply and reshape it, which can done as follows

```
def transpose_convolution(x, filter):
    W = matrix_transposed(filter)
    y = torch.mv(W, x.reshape(-1))
    return y.reshape((3,3))
```