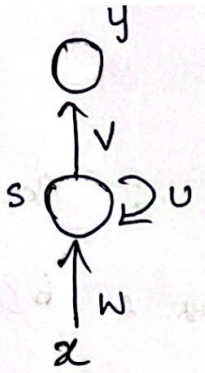
1)



Given bias = 0

hidden state (h)

$h = \tanh(Wx + Us)$

$y = \text{sigmoid}(Vh)$

$$W = \begin{bmatrix} 0.5 & 0 & 0.1 \\ 0.1 & -0.2 & 0.2 \end{bmatrix}$$

$$U = \begin{bmatrix} 1 & 0 \\ 0.1 & 0.5 \end{bmatrix}$$

$$V = \begin{bmatrix} 1.1 & 0.8 \end{bmatrix}$$

→ step 1 [input 1]

Forward Pass :-

$$x = \begin{bmatrix} 1 \\ 5 \\ 0.1 \end{bmatrix}$$

$$s = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$h = \tanh\left(\begin{bmatrix} 0.5 & 0 & 0.1 \\ 0.1 & -0.2 & 0.2 \end{bmatrix} \begin{bmatrix} 1 \\ 5 \\ 0.1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0.1 & 0.5 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right)$$

$$= \tanh\left(\begin{bmatrix} 0.5+0+0.01 \\ 0.1-1+0.02 \end{bmatrix} + \begin{bmatrix} 0+0 \\ -0.5 \end{bmatrix} \right)$$

$$= \tanh\left(\begin{bmatrix} 0.51 \\ -0.88 \end{bmatrix} + \begin{bmatrix} 0 \\ -0.5 \end{bmatrix} \right)$$

$$= \tanh\left(\begin{bmatrix} 0.51 \\ -1.38 \end{bmatrix} \right) = \begin{bmatrix} 0.47 \\ -0.88 \end{bmatrix}$$

$$\hat{y}_{act} = \sigma\left(\begin{bmatrix} 1.1 & 0.8 \end{bmatrix} \begin{bmatrix} 0.47 \\ -0.88 \end{bmatrix} \right)$$

$$\hat{y}_{act} = \sigma(1.1 \times 0.47 + 0.8 \times -0.88)$$

$$\hat{y}_{act} = \frac{1}{1 + e^{+0.187}}$$

$$= 0.45$$

Since $\hat{y}_{act} < 0.5 \Rightarrow \hat{y} = 0$

error =

step2 [Input 2]

$$x = \begin{bmatrix} 10 \\ -2 \\ 0.5 \end{bmatrix}$$

$$h = \tanh\left(\begin{bmatrix} 0.5 & 0 & 0.1 \\ 0.1 & -0.2 & 0.2 \end{bmatrix}\begin{bmatrix} 20 \\ -2 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0.1 & 0.5 \end{bmatrix}\begin{bmatrix} 0.47 \\ -0.88 \end{bmatrix}\right)$$

$$h = \tanh\left(\begin{bmatrix} 10+0+0.05 \\ 2+0.4+0.1 \end{bmatrix} + \begin{bmatrix} 0.47+0 \\ 0.047-0.44 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 10.05 \\ 2.5 \end{bmatrix} + \begin{bmatrix} 0.47 \\ -0.393 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 10.52 \\ 2.107 \end{bmatrix}\right)$$

$$= \begin{bmatrix} 1 \\ 0.97 \end{bmatrix}$$

$$\hat{y}_{act} = \sigma\left(\begin{bmatrix} 1.1 & 0.8 \end{bmatrix}\begin{bmatrix} 1 \\ 0.97 \end{bmatrix}\right) = \sigma(1.1 + 0.8 \times 0.97)$$

$$= \frac{1}{1+e^{-1.876}}$$

$$= 0.87$$

Since $\hat{y}_{act} > 0.5 \rightarrow \hat{y} = 1$

step3 [Input 3]

$$x = \begin{bmatrix} 0 \\ 3 \\ 0.2 \end{bmatrix}$$

$$h = \tanh\left(\begin{bmatrix} 0.5 & 0 & 0.1 \\ 0.1 & -0.2 & 0.2 \end{bmatrix}\begin{bmatrix} 0 \\ 3 \\ 0.2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0.1 & 0.5 \end{bmatrix}\begin{bmatrix} 1 \\ 0.97 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 0+0+0.02 \\ 0-0.6+0.04 \end{bmatrix} + \begin{bmatrix} 0.97+0 \\ 0.1+0.485 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 0.02 \\ -0.56 \end{bmatrix} + \begin{bmatrix} 0.97 \\ 0.585 \end{bmatrix}\right) = \tanh\left(\begin{bmatrix} 0.99 \\ 0.025 \end{bmatrix}\right)$$

$$h = \begin{bmatrix} 0.76 \\ 0.025 \end{bmatrix}$$

$$\hat{y}_{act} = \sigma\left(\begin{bmatrix} 1.1 & 0.8 \end{bmatrix}\begin{bmatrix} 0.76 \\ 0.025 \end{bmatrix}\right) = \sigma\left(1.1 \times 0.76 + 0.8 \times 0.025\right)$$

$$= \frac{1}{1 + e^{-0.856}}$$

$$= 0.70$$

Since $\hat{y}_{act} > 0.5 \Rightarrow \hat{y} = 1$

Predict probabilities for 3 inputs are

$$[0.45, 0.87, 0.70]$$

Predictions are $[0, 1, 1] = [\text{Study}, \text{Skiing}, \text{Skiing}]$

② $A = \begin{bmatrix} 1 & 0 & -1 & 0 \end{bmatrix}$ $\quad Q = \begin{bmatrix} 0 & 3 \\ 1 & 2 \\ -1 & 0 \\ 1 & 1 \end{bmatrix}$ $\quad K = \begin{bmatrix} -0.5 & 1 \\ 1 & 0.5 \\ 0 & 0 \\ -1 & 1 \end{bmatrix}$

$B = \begin{bmatrix} 0 & 2 & 0.5 & -1 \end{bmatrix}$ $\quad V = \begin{bmatrix} 1 & 1 \\ -1 & 2 \\ 3 & -1 \\ 2 & 1 \end{bmatrix}$

$$Q_A = A \times Q = \begin{bmatrix} 1 & 0 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 3 \\ 1 & 2 \\ -1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \end{bmatrix}$$

$$K_A = A \times K = \begin{bmatrix} 1 & 0 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} -0.5 & 1 \\ 1 & 0.5 \\ 0 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} -0.5 & 1 \end{bmatrix}$$

$$V_A = A \times V = \begin{bmatrix} 1 & 0 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ -1 & 2 \\ 3 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 2 \end{bmatrix}$$

$$Q_B = B \times Q = \begin{bmatrix} 0 & 2 & 0.5 & -1 \end{bmatrix} \times \begin{bmatrix} 0 & 3 \\ 1 & 2 \\ -1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.5 & 3 \end{bmatrix}$$

$$K_B = B \times K = \begin{bmatrix} 0 & 2 & 0.5 & -1 \end{bmatrix} \times \begin{bmatrix} -0.5 & 1 \\ 1 & 0.5 \\ 0 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \end{bmatrix}$$

$$V_B = B \times V = \begin{bmatrix} 0 & 2 & 0.5 & -1 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ -1 & 2 \\ 3 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} -2.5 & 2.5 \end{bmatrix}$$

Embedding of A $= (\theta_A \cdot K_A) V_A + (\theta_A \cdot K_B) V_B$

$$= \left(\begin{bmatrix} 1 & 3 \end{bmatrix} \times \begin{bmatrix} -0.5 \\ 1 \end{bmatrix} \right) \begin{bmatrix} -2 & 2 \end{bmatrix} +$$

$$\left(\begin{bmatrix} 1 & 3 \end{bmatrix} \times \begin{bmatrix} 3 \\ 0 \end{bmatrix} \right) \begin{bmatrix} -2.5 & 2.5 \end{bmatrix}$$

$$= \frac{e^{2.5}}{e^{2.5}+e^{3}} \begin{bmatrix} -2 & 2 \end{bmatrix} + \frac{e^{3}}{e^{2.5}+e^{3}} \begin{bmatrix} -2.5 & 2.5 \end{bmatrix}$$

$$= \cancel{[-12.5 \quad 12.5]} = [-2.31 \quad 2.31]$$

Embedding of B $= (\theta_B \cdot K_A) V_A + (\theta_B \cdot K_B) V_B$

$$= \left(\begin{bmatrix} 0.5 & 3 \end{bmatrix} \times \begin{bmatrix} -0.5 \\ 1 \end{bmatrix} \right) \begin{bmatrix} -2 & 2 \end{bmatrix} +$$

$$\left(\begin{bmatrix} 0.5 & 3 \end{bmatrix} \times \begin{bmatrix} 3 \\ 0 \end{bmatrix} \right) \begin{bmatrix} -2.5 & 2.5 \end{bmatrix}$$

$$= \frac{e^{2.75}}{e^{2.75}+e^{1.5}} \begin{bmatrix} -2 & 2 \end{bmatrix} + \frac{e^{1.5}}{e^{1.5}+e^{2.75}} \begin{bmatrix} -2.5 & 2.5 \end{bmatrix}$$

$$= \cancel{[-9.25 \quad 9.25]}$$

$$= \cancel{[-2.95 \quad 2.05]}$$

$$= [-2.11 \quad 2.11]$$

3) a) Multi headed attention enables the model to focus on and learn from different tokens of the sentence rather than focusing on a single token alone. It allows capturing various representations and dependencies on other tokens. By aggregating information from multiple heads, the model can potentially capture richer and more nuanced features. Multi-headed attention also provides a way to regularize the model by learning different sets of attention weights, which can help reduce overfitting by providing different perspectives on the input data.

b) If a transformer has an attention layer with five transformer heads, then it should have five key matrices.

4)

RNN's	**Transformers**
Look at a single token/ word at a time.	Consider all tokens/words in getting embedding.
It is sequential in nature, order of tokens matters.	Sequence of tokens doesn't matter.
They may suffer from the vanishing or exploding gradient problem during training, limiting their ability to retain information over long sequences.	They can create a dynamic memory that can adapt to different inputs without suffering from the same vanishing gradient issues.