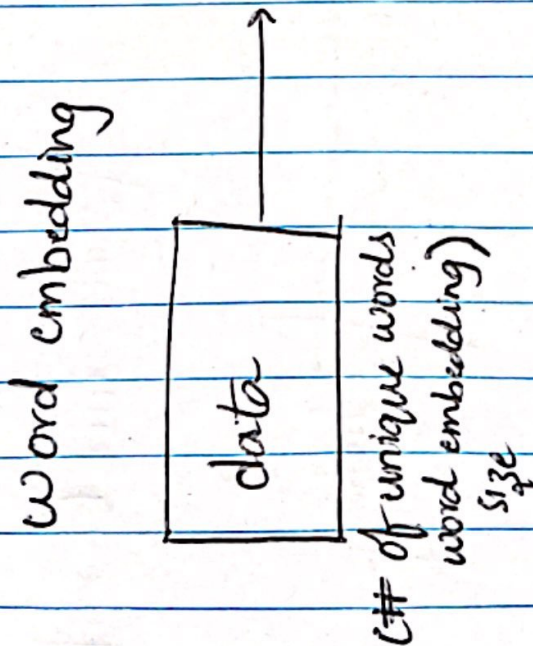
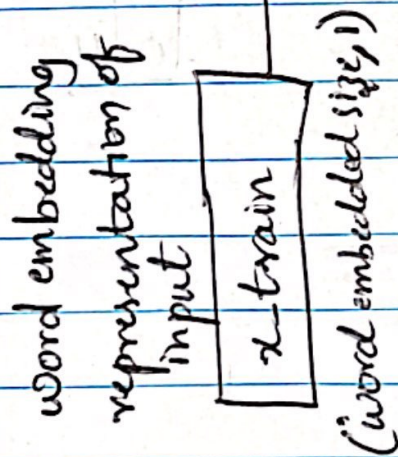
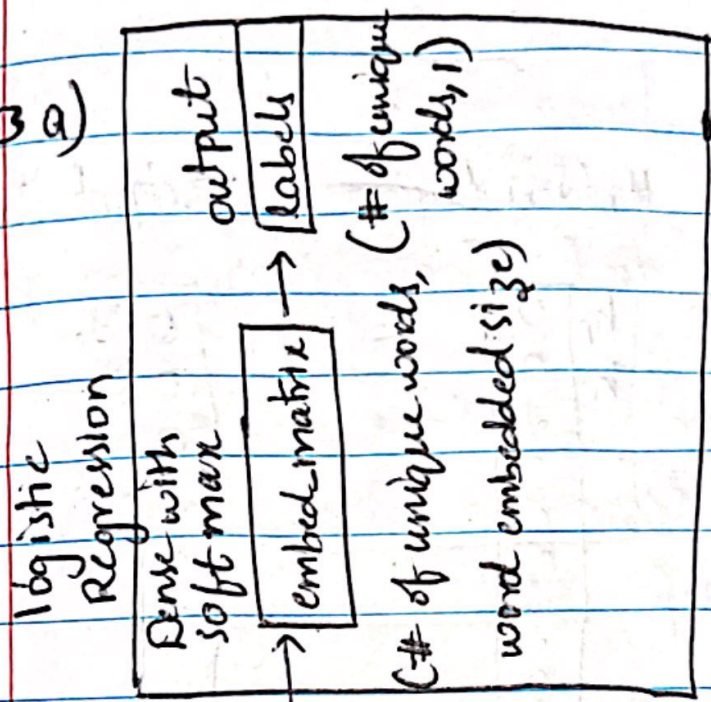


prob-3 a)



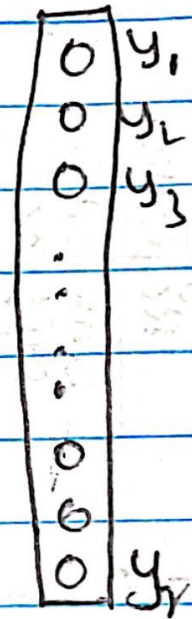
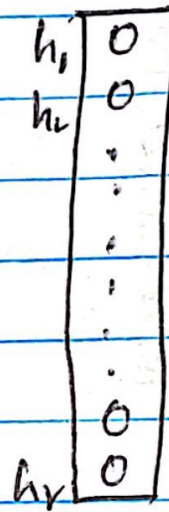
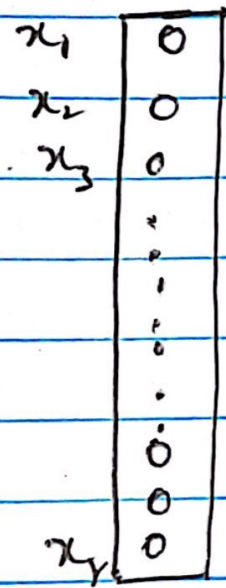
3)

b, c)

Input layer

Hidden layer

output layer





prob-9)

a)

$$K = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \\ 0 & 0 & 10 \end{bmatrix}$$

$$V = \begin{bmatrix} 1 & 0 \\ 10 & 0 \\ 100 & 5 \\ 1000 & 6 \end{bmatrix}$$

$$Q = \begin{bmatrix} 0 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 10 & 0 \end{bmatrix}$$

$$QK^T = \begin{bmatrix} 0 & 0 & 100 & 100 \\ 0 & 100 & 0 & 0 \\ 100 & 100 & 0 & 0 \end{bmatrix}$$

$$QK^TV = \begin{bmatrix} 0 & 0 & 100 & 100 \\ 0 & 100 & 0 & 0 \\ 100 & 100 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 10 & 0 \\ 100 & 5 \\ 1000 & 6 \end{bmatrix}$$

$$= \begin{bmatrix} 11000 & 1100 \\ 1000 & 0 \\ 1100 & 0 \end{bmatrix}$$

$$\text{Softmax}(QK^TV) = \begin{bmatrix} \sim 1 & \sim 0 \\ \sim 1 & \sim 0 \\ \sim 1 & \sim 0 \end{bmatrix}$$

b) The attention function is used to solve the bottleneck problem in recurrent neural networks and the vanishing gradient problem.

The softmax function is used to convert the score to probabilities so that they sum up to 1.

$$\text{softmax}(x) = \frac{e^{x_i}}{\sum_x e^{x_i}}$$