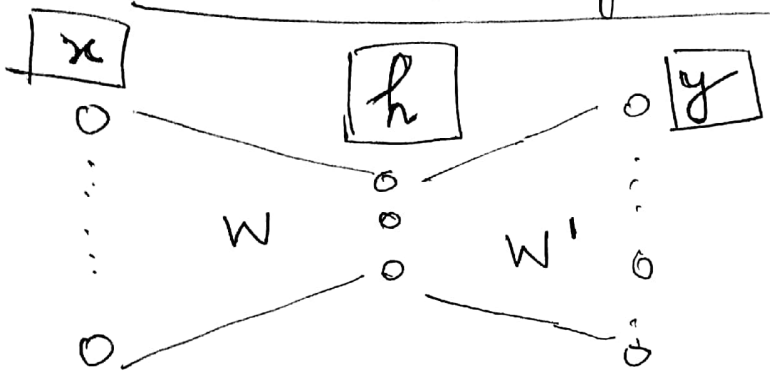
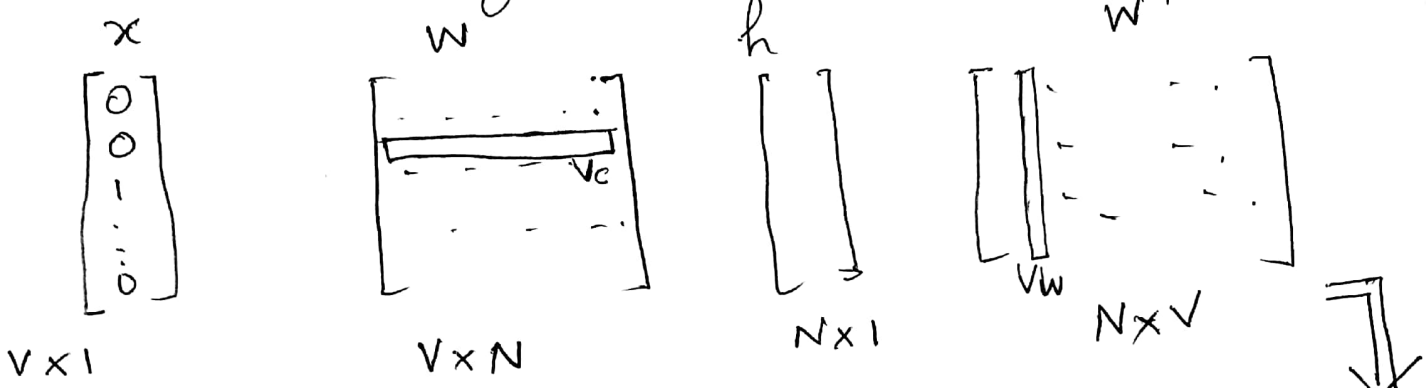
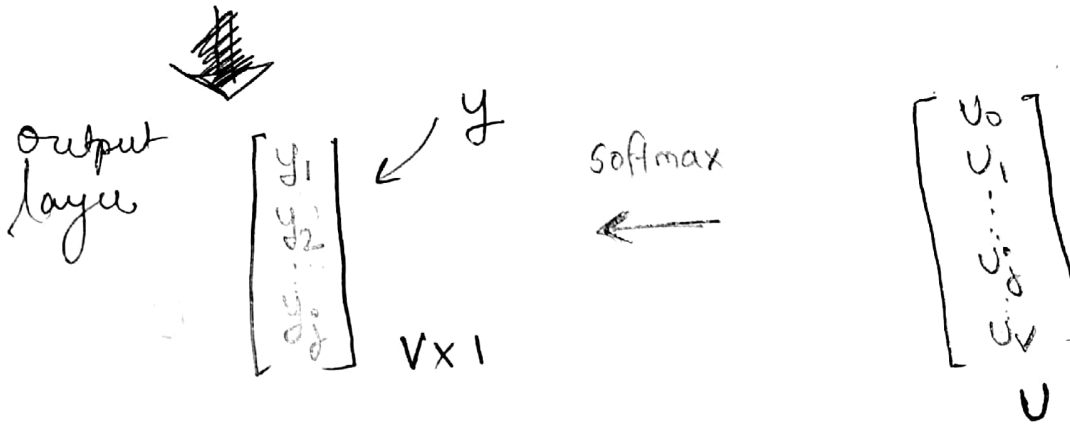
Word Embeddings - Skip Gram model.



x - one hot representation of word w (the context word)

h - hidden layer.





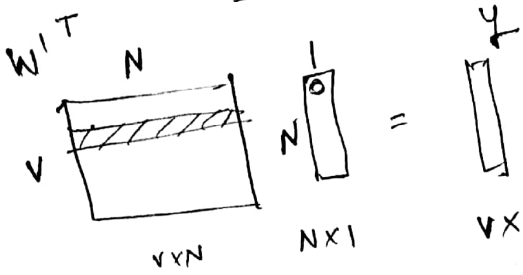
$h = x^T W = W^T x$ { $x^T w$ is a row of matrix W so h is a row of W (a row of context w

$$\boxed{h = v_c = W^T x}$$ { ∵ $W^T x$ is of N×1 = size

$W'^T h \Rightarrow U$ (output layer)

$V \times N \quad N \times 1 \Rightarrow V \times 1$

so $\boxed{U = W'^T h}$



Let V_w^T be a row in W'^T, it means V_w is a row in W'^T. so if we multiply V_w^T with V_c we get our output (U). Now we apply softwax to this 'U' to get y. In V_w, w (subscript) means any word w in our corpus.

Let U_i be an neuron (i^{th} neuron) in output layer.

$$\boxed{U_i = V_w^T \cdot V_c}$$

$$y_i = \sigma(U_i) = \frac{e^{U_i}}{\sum_j e^{U_j}} = \frac{e^{V_w^T V_c}}{\sum_{w'} e^{V_{w'}^T V_c}}$$

w' means any word w'

and $\sum_{w'}$ means summation over all words.

we can also write $\boxed{y_i = \frac{e^{V_c^T V_w}}{\sum_{w'} e^{V_c^T V_{w'}}}}$

$y_i = P(w|c)$ = prob. of a word w given that a context word c is given

Our parameters are $= \{v_c, v_w\}$

$\theta = \{v_c, v_w\}$

$$L(\theta) = \prod P(w|c;\theta) \quad \{\text{Product of probabilities}\}$$

$$= \prod \frac{e^{v_c^T v_w}}{\sum_{w'} e^{v_c^T v_{w'}}}$$

To maximize this we can take Maximum Log Likelihood

$$\ell(\theta) = \sum_{w'} v_c^T v_w - \log \sum_{w'} e^{v_c^T v_{w'}}$$

Let find $\frac{\partial \ell}{\partial v_w}$ first

$$\frac{\partial \ell}{\partial v_w} = v_c - \frac{1}{\sum_{w'} e^{v_c^T v_{w'}}} \times e^{v_c^T v_{w'}} \times v_c$$

$$= v_c - P(w|c)\, v_c$$

$$\boxed{\frac{\partial \ell}{\partial v_w} = v_c \left(1 - P(w|c)\right)}$$

Now we can apply gradient descent to find optimal parameters

Gradient Descent :

$$V_w^{new} \leftarrow V_w^{old} - \alpha \cdot V_c \left(1 - P(w|c)\right)$$

α ↓ learning rate