CBOW MODEL, ONE WORD CONTEXT $C=1$ → CONTRAST SKIP GRAM

INPUT VECTOR: 1-HOT ENCODING OF LENGTH $V$

(PUTS ALL PROB MASS ON the 1 WORD)

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ \vdots \\ x_V \end{bmatrix}$$

FOR INPUT WORD $k$,
$x_k = 1$ AND $\forall j, j \neq k \; x_j = 0$

$W = V \times N$ ($N$ HIDDEN NODES/UNITS); INPUT → HIDDEN LAYER

EACH ROW OF $W$ AN $N$-DIMENSIONAL VECTOR

REPRESENTATION OF INPUT WORD $w$, $v_w$

HIDDEN LAYER $h$

$h = x^T W = W_k := \vec{V}_{W_I}$

$\vec{V}_{W_I}$ IS THE VECTOR REPRESENTATION OF INPUT WORD $W_I$

$$h = \underbrace{\begin{bmatrix} 0 & \cdots & \overset{k}{1} & \cdots & 0 \end{bmatrix}}_{x^T : 1 \times V} \overbrace{\begin{bmatrix} V_1 \\ | \\ | \\ V_V \end{bmatrix}}^{W} = \begin{bmatrix} 0 & \cdots & 1 & \cdots \end{bmatrix}_{1 \times N} = h$$

$V \times N$

$k^{th}$ ROW OF $W$

$1 \times N$

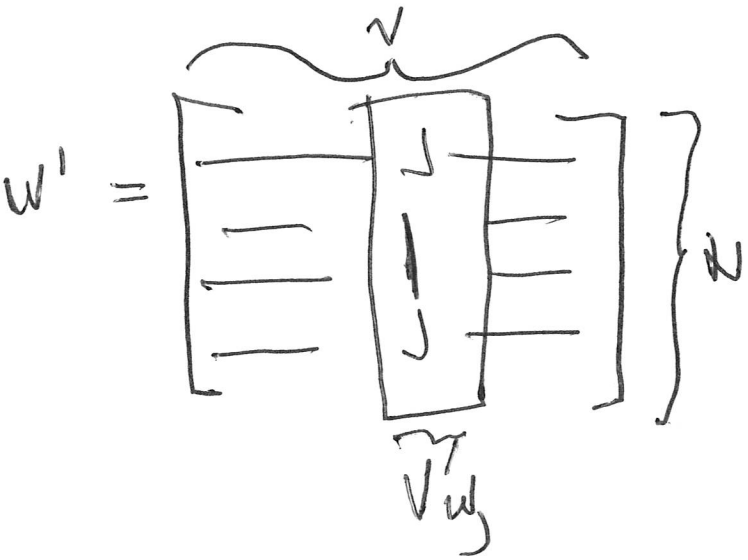
$x^T W = h \Rightarrow$ COPIES THE $k^{th}$ ROW OF $W$ TO $h$

HIDDEN LAYER → OUTPUT LAYER

$W' = \{w'_{ij}\}$ $N \times V$

COMPUTE SCORE $u_j$

$u_j = v'^{T}_{w_j} \cdot h$ $\quad v'_{w_j}$ IS THE $j^{th}$ column of $W'$




$$v'^{T}_{w_j} \cdot h = \underbrace{\begin{bmatrix} - & h & - \end{bmatrix}}_{1 \times V} \cdot \underbrace{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{1 \times h} = \text{SCALAR} = u_j$$

$v'^{T}_{w_j}$ $\qquad$ $h$

GENERAL SOFTMAX

$$P(w_j \mid w_I) = y_j = \frac{\exp(u_j)}{\sum_{j'=1}^{V} \exp(u_{j'})}$$

$y_j$: output of $j^{th}$ node in output layer

$u_j = v'^{T}_{w_j} \cdot h$

$= v'^{T}_{w_j} \cdot x^T W$

$$p_\theta(y \mid x) = \frac{\exp(\theta^T f(x,y))}{\sum_{y' \in Y(x)} \exp(\theta^T f(x,y'))}$$

$h = x^T W := V_{W_I}$ (VECTOR REP of INPUT WORD $W_I$)

$u_j = V'^T_{w_j} \cdot h$ ← $x^T W := V_{W_I}$

↑ SCORE $j^{th}$ column of $W'$ $(N \times V)$ $\Rightarrow \begin{bmatrix} w'_{1,j} \\ \vdots \\ w'_{N,j} \end{bmatrix}$

SOFTMAX POSTERIOR DISTRIBUTION

$$p(w_j | w_I) = y_j = \frac{\exp(u_j)}{\sum_{j'=1}^{V} \exp(u_{j'})}$$

↑ OUTPUT OF $j^{th}$ NODE IN OUTPUT LAYER

$u_j = V'^T_{w_j} \cdot h$

ISSUE: DOT PRODUCT? WRITTEN LIKE OUTER PRODUCT ?

$$p(w_j | w_I) = \frac{\exp(V'^T_{w_O} V_{W_I})}{\sum_{j'=1}^{V} \exp(V'^T_{w_{j'}} V_{W_I})}$$

NOTE: $V_W$ AND $V_{W'}$ ARE 2 REPRESENTATIONS OF INPUT WORD $W$

$V_W$ COMES FROM $W$, the INPUT → HIDDEN WEIGHT MATRIX

↑ the rows of

$V_{W'}$ – columns of $W'$, HIDDEN → OUTPUT

# TRAINING

IDEA IS TO MAXIMIZE

$$p(w_j \mid w_I) = \frac{\exp(v'^{T}_{w_O} v_{w_I})}{\sum_{j'=1}^{V} \exp(v'^{T}_{w_{j'}} v_{w_I})}$$

$$\Rightarrow \max(p(w_O \mid w_I)) = \max y_{j^*}$$

$$= \max \log y_{j^*}$$

$$= \max \log \left( \frac{\exp(u_{j^*})}{\sum_{j'=1}^{V} \exp(u_{j'})} \right)$$

$$= \max \log(\exp(u_{j^*})) - \log \sum_{j'=1}^{V} \exp(u_{j'})$$

$$= u_{j^*} - \log \sum_{j'=1}^{V} \exp(u_{j'})$$

$$:= -E$$

$$= -v'^{T}_{w_O} \cdot h + \log \sum_{j'=1}^{V} \exp(v'^{T}_{w_{j'}} \cdot h)$$

$$u_{j^*} - \log \sum_{j'=1}^{V} \exp(u_{j'})$$

Log

SO $E = -\log p(w_O \mid w_I)$ IS THE LOSS FUNCTION (WANT TO MINIMIZE THIS), $j^*$ IS THE INDEX OF THE ACTUAL OUTPUT WORD IN THE OUTPUT LAYER

BTW, $E = -\log p(w_O | w_T)$ IS A SPECIAL CASE OF THE CROSS ENTROPY ($\hat{y}=1$)

UPDATE HIDDEN → OUTPUT

NEED THE DERIVATE OF E WRT the $j^{th}$ NODE'S INPUT, NAMELY $u_j$ ⇒

$$\frac{\partial E}{\partial u_j} = \partial \left( u_{j^*} - \log \left( \sum_{j'=1}^{V} \exp(u_{j'}) \right) \right)$$

$$= \frac{\partial u_{j^*}}{\partial u_j} - \frac{\partial \log \sum_{j'=1}^{V} \exp(u_{j'})}{\partial u_j}$$

$$= t_j - \frac{1}{\sum_{j'=}^{V} \exp(u_{j'})} \quad \text{where}$$

$t_j = 1(j = j^*)$ ie 1 if $j = j^*$, 0 otherwise

$$\frac{1}{\sum_{j'=1}^{V} \exp(u_{j'})} \sim y_j$$

(ISSUE - REALLY NEED $\exp(u_j)$ in the NUMERATOR) *

$$\Rightarrow \frac{\partial E}{\partial u_j} = y_i - t_i := e_j$$ ← PREDICTION ERROR $e_j$ in the OUTPUT LAYER

NEXT: DERIVATE OF $w'_{ij}$ TO GET

The gradient of the HIDDEN→OUTPUT WEIGHTS

$$\frac{\partial E}{\partial w'_{ij}} = \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial w'_{ij}} \quad \text{(chain rule)}$$

$$= e_j \cdot \frac{\partial u_j}{\partial w'_{ij}}$$

$$= e_j \cdot \frac{\partial (V_{w'_j}^T \cdot h)}{\partial w'_{ij}} \qquad V_{w'_j} = \begin{bmatrix} w'_{1j} \\ w'_{ij} \\ w'_{Nj} \end{bmatrix}$$

$$= e_j \cdot h_i \qquad V^T_{w'_j} = \begin{bmatrix} w'_{1j} & \cdots & w'_{ij} & \cdots & w'_{Nj} \end{bmatrix}$$

---

UPDATE INPUT → HIDDEN

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{V} \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial h_i} \quad \text{(chain rule)}$$

$$= \sum_{j=1}^{V} e_j \cdot w'_{ij} := EH_i$$

WHAT EXACTLY IS $w'_{ij}$ ?

$$\frac{\partial V^T_{w'_j}}{\partial w_{ij}} = [0 \cdots 1 \cdots 0] \quad \text{(1 in the } i\text{th)}$$

$$\frac{\partial h}{\partial w'_{ij}} = \vec{h} \Rightarrow$$

$$\frac{\partial (V^T_{w'_j} \cdot \vec{h})}{\partial w_{ij}} = h_i$$

← SELECTS the $i$th ELT of $h$