a

$$-\sum_{w \in Voc} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$$

$$-\sum_{w \in Voc} y_w \log(\hat{y}_w) \overset{\substack{w \neq o \\ y_w = 0}}{=} -y_o \log(\hat{y}_o) = -\log(\hat{y}_o)$$

(b) $$\frac{\partial J(v_c, o, U)}{\partial v_c} = ?$$

$$= \frac{\partial - \log \frac{e^{u_o^T v_c}}{\sum_w e^{u_w^T v_c}}}{\partial v_c}$$

$$= \frac{-\partial u_o^T v_c}{\partial v_c} + \frac{\partial \log \sum_w e^{u_w^T v_c}}{\partial v_c}$$

$$= -u_o + \frac{\sum_t e^{u_t^T v_c} u_t}{\sum_w e^{u_w^T v_c}}$$

$$= -u_o + \sum_t \frac{e^{u_t^T v_c} u_t}{\sum_w e^{u_w^T v_c}}$$

$$= -u_o + \sum_t \hat{y}_t u_t$$

$$= -Uy + U\hat{y} = U(\hat{y} - y)$$

C

$w = o$

$$\frac{\partial\left(-u_o^T v_c + \log \sum_w e^{u_w^T v_c}\right)}{\partial u_o}$$

$$= -v_c + \frac{e^{u_o^T v_c}\, v_c}{\sum_t e^{u_t^T v_c}}$$

$$= -v_c + \hat{y}_o v_c$$

$w \neq o$

$$\frac{\partial\left(-u_o^T v_c + \log \sum_w e^{u_w^T v_c}\right)}{\partial u_w}$$

$$= 0 + \frac{e^{u_w^T v_c}\, v_c}{\sum_t e^{u_t^T v_c}}$$

$$= \hat{y}_w v_c$$

d

$U : [D \times |W|]$

$$v_c' = v_c \begin{array}{c} \scriptstyle 0^{th}\text{ index} \\ \scriptstyle \downarrow \end{array}\!\!\!\!\!\!\!\!\!\!\!\! \left[0\ 0\ 0 \dots -1 \dots 0\ 0\ 0\right]_{[1 \times |W|]}$$

$$\frac{\partial J}{\partial U} = v_c' + \underset{[D \times 1]}{v_c}\ \underset{[1 \times |W|]}{\hat{y}^T}$$

(e)

$$\left(\frac{e^x}{e^x+1}\right)' = \frac{e^x(e^x+1)-(e^x)(e^x)}{(e^x+1)^2} = \frac{e^x(e^x+1-e^x)}{(e^x+1)(e^x+1)}$$

$$= \frac{e^x}{e^x+1} \times \frac{1}{e^x+1} = \sigma(x)(1-\sigma(x))$$

f

$$\frac{\partial\left(-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K}\log(\sigma(-u_k^T v_c))\right)}{\partial v_c}$$

$$= \frac{-\cancel{\sigma(u_o^T v_c)}(1-\sigma(u_o^T v_c))u_o}{\cancel{\sigma(u_o^T v_c)}} - \sum_{k=1}^{K} \frac{\cancel{\sigma(-u_k^T v_c)}(1-\sigma(-u_k^T v_c))(-u_k)}{\cancel{\sigma(-u_k^T v_c)}}$$

$$= -(1-\sigma(u_o^T v_c))u_o + \sum_{k=1}^{K} (1-\sigma(-u_k^T v_c))u_k$$

$$\frac{\partial\left(-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K}\log(\sigma(-u_k^T v_c))\right)}{\partial u_o}$$

$$= \frac{-\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))v_c}{\sigma(u_o^T v_c)} - 0$$

$$= -(1-\sigma(u_o^T v_c))v_c$$

$$\frac{\partial\left(-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))\right)}{\partial u_k}$$

$$= 0 - \frac{\sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c))(-v_c)}{\sigma(-u_k^T v_c)}$$

$$= (1-\sigma(-u_k^T v_c))v_c$$

more efficient because this approach computes over K samples ($O(K)$) but in naive softmax computing the denominator costs $O(|V|)$ and we know that $K \ll |V|$

(g) suppose that $\underline{s}$ samples are equal to u_k

$$\frac{\partial \left(-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))\right)}{\partial u_k}$$

$$= 0 - (s)\left(\frac{\sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c))(-v_c)}{\sigma(-u_k^T v_c)}\right) + 0$$

$$= s(1-\sigma(-u_k^T v_c))v_c$$

(h)

(i) $$\sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J_{skip\text{-}gram}(v_c, w_{t+j}, U)}{\partial U}$$

(ii) $$\sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J_{skip\text{-}gram}(v_c, w_{t+j}, U)}{\partial v_c}$$

(iii) 0

One thing that can easily be seen is analogy of king : male :: queen : female. Also proximity of similar words such as “female” and “woman” is visible.