outside u_w center v_w

$U = [\ \ ]$ $V = [\ \ ]$ $w \in Vocab$

$$P(O=o|C=c) = \frac{e^{u_o^T v_c}}{\sum_{w \in Voc} e^{u_w^T v_c}};$$

$J_{naive\text{-}softmax} = -\log P(O=o|C=c)$

$J_{cross\text{-}ent} = -\sum y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$; y = one hot vector,

$$\hat{y} = P(O|C=c) = \frac{e^{u_o^T v_c}}{\sum_w e^{u_w^T v_c}}$$

$$J_{naive\text{-}softmax} = -\log\left[\frac{e^{u_o^T v_c}}{\sum_{w \in Voc} e^{u_w^T v_c}}\right] = -u_o^T v_c + \log \sum_{w \in V} e^{u_w^T v_c}$$

$$J_{cross\text{-}ent} = -\log \hat{y}_o = -\log\left[\frac{e^{u_o^T v_c}}{\sum_{w \in V} e^{u_w^T v_c}}\right]$$

a) $J_{naive\text{-}softmax} = -\log P(O=o|C=c) = -\log(\hat{y}_o) = -\sum_{w \in V} y_w \log(\hat{y}_w) = J_{cross\text{-}entropy}$

Since that y_w is hot vector and will 1 if $O=o$ we get $-\log(\hat{y}_o)$, and otherwise 0. while $-\log(\hat{y}_o)$ is rewrite format of $-\log(P(O=o|C=c))$.

b) $\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}(-\log(\hat{y}_o)) = \frac{\partial}{\partial v_c}\left(-u_o^T v_c + \log \sum_{w \in V} e^{u_w^T v_c}\right) = -u_o + \frac{1}{\sum_{w \in V} e^{u_w^T v_c}} \cdot \sum_{w \in V} e^{u_w^T v_c} \cdot u_w$

$= -u_o + \sum_{w \in V} \frac{e^{u_w^T v_c}}{\sum_{w \in V} e^{u_w^T v_c}} \cdot u_w = -u_o + \sum_{w \in V} P(O=w|C=c) \cdot u_w = -\sum_{w \in V} y_w u_w + \sum \hat{y}_w u_w$

$P(O=w|C=c)$

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

1) $\frac{\partial J}{\partial v_c} = 0$, if $\hat{y} = y$ when predicted outside word is true outside word.

2) because its adjust v_c which decrease loss.

c) $\frac{\partial J}{\partial u_w} = \frac{\partial}{\partial u_w}\left(-u_o^T v_c + \log \sum_{w \in V} e^{u_w^T v_c}\right) = \frac{\partial}{\partial u_w}(-u_o^T v_c) + \sum_{w \in V} \frac{e^{u_w^T v_c}}{\sum_{w \in V} e^{u_w v_c}} \cdot v_c$

if $w = o$: $\frac{\partial J}{\partial u_w} = -v_c + P(O=o|C=c) \cdot v_c = -v_c + \hat{y} v_c$

if $w \neq o$ $\frac{\partial J}{\partial u_w} = 0 + P(O=o|C=c) v_c = \hat{y} v_c$

$\Rightarrow \frac{\partial J}{\partial u_w} = (\hat{y}_w - y_w) v_c$ for $w = 1 \dots |V|$

d) $\frac{\partial J}{\partial U} = \left[\frac{\partial J}{\partial u_1}, \frac{\partial J}{\partial u_2}, \dots, \frac{\partial J}{\partial u_{|V|}}\right] = [\hat{y}_1 v_c, \hat{y}_2 v_c, \dots, -v_c + \hat{y} v_c, \dots, \hat{y}_{|V|} v_c] = (\hat{y} - y_o) v_c^T$

e) $f(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$ $\frac{\partial f}{\partial x} = \begin{cases} 1 & x > 0 \\ 0 & x < 0 \end{cases}$

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

$\frac{\partial(\sigma(x))}{\partial x} = \frac{(e^x)'(e^x+1) - e^x(e^x+1)'}{(e^x+1)^2} = \frac{e^{2x} + e^x - e^{2x}}{(e^x+1)^2} = \frac{e^x}{(e^x+1)} \cdot \frac{1}{e^x+1}$

$e^x \cdot \sigma(x) + \sigma(x) = e^x$

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

$= \sigma(x) \cdot \frac{1}{\frac{\sigma(x)}{1-\sigma(x)} + 1} = \sigma(x)(1 - \sigma(x))$

g) $J_{\text{neg-sample}}(v_c, o, \bar{U}) = -\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K} \log(\sigma(-u_{ws}^T v_c))$

gb) $\frac{\partial J_{\text{neg-sam}}}{\partial v_c} = \frac{\partial[-\log \sigma(u_o^T v_c)]}{\partial v_c} - \frac{\partial}{\partial v_c}\left[\sum_{s=1}^{K} \log(\sigma(-u_{ws}^T v_c)\right]$

(i)

$\cdot = u_{ws}^T v_c$

g b) $\frac{\partial J_{n-s}}{\partial v_c} = -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c)(1-\sigma(u_o^T v_c)) \cdot u_o - \sum_{s=1}^{K} \frac{1}{\sigma(\cdot)} \cdot \sigma(\cdot)(1-\sigma(\cdot))(-u_{ws})$

$\frac{\partial J_{n-s}}{\partial v_c} = -(1-\sigma(u_o^T v_c)) \cdot u_o + \sum_{s=1}^{K} (1-\sigma(-u_{ws}^T v_c)) \cdot u_{ws} \cdot (-1)$

g c) $\frac{\partial J_{n-s}}{\partial u_o} = -\frac{\partial [\log \sigma(u_o^T v_c)]}{\partial u_o} - \underbrace{\frac{\partial}{\partial u_o}\left[\sum_{s=1}^{K} \log(\sigma(-u_{ws}^T v_c))\right]}_{0}$

$\frac{\partial J_{n-s}}{\partial u_o} = -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c) \cdot (1-\sigma(u_o^T v_c)) \cdot v_c = -(1-\sigma(u_o^T v_c)) v_c$

g 3) $\frac{\partial J_{n-s}}{\partial u_{ws}} = -\underbrace{\frac{\partial [\log \sigma(u_o^T v_c)]}{\partial u_{ws}}}_{0} - \frac{\partial}{\partial u_{ws}}\left[\sum_{s=1}^{K} \log[\sigma(-u_{ws}^T v_c)]\right]$

$\frac{\partial J_{n-s}}{\partial u_{ws}} = -\frac{1}{\sigma(-u_{ws}^T v_c)} \cdot \sigma(-u_{ws}^T v_c) \cdot (1-\sigma(-u_{ws}^T v_c)) \cdot (-v_c)$

$\frac{\partial J_{n-s}}{\partial u_{ws}} = +(1-\sigma(-u_{ws}^T v_c)) \cdot v_c$

(ii) to aviod repetition we can use $[-(1-\sigma(u_o^T v_c))v_c]$ this expersion since that it is used in the two formulas. If I write in terms of V it is u_o

(iii) It is efficently use memory with fixed size K. While naive-softmax use whole outside words.

h) $\frac{\partial J_{neg\text{-}sa}}{\partial u_{ws}} = -\frac{\partial}{\partial u_{ws}} \sum_{s=1}^{K} \log(\sigma(-u_{ws}^T v_s)) + 0 = -\frac{\partial}{\partial u_{ws}} \sum_{\substack{u_{wk}=u_{ws} \\ k=1,\dots,K}}^{K} \log(\sigma(-u_{wk}^T v_c)$

$= -\frac{\partial}{\partial u_{ws}} \sum_{\substack{u_{wk}\neq u_{ws} \\ k=1,\dots,K}}^{K} \log(\sigma(-u_{wk}^T v_c)) =$

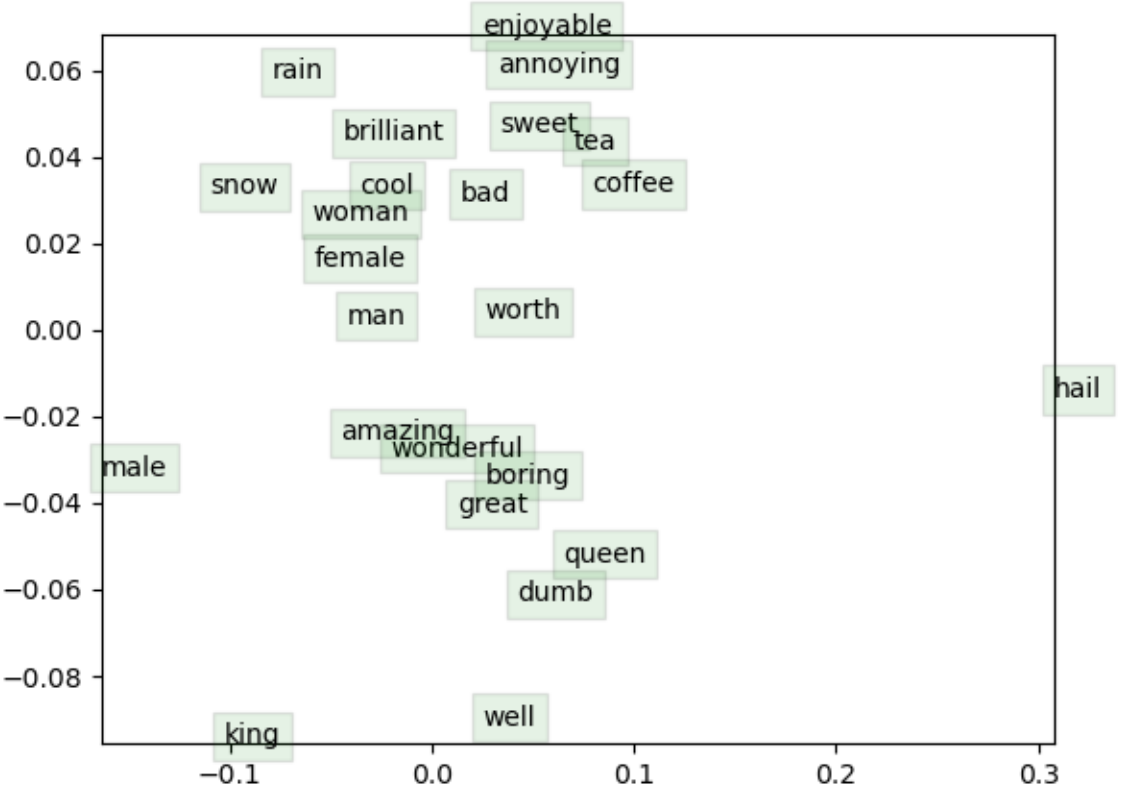
$\frac{\partial J_{neg}}{\partial u_{ws}} = -\sum_{\substack{u_{wk}=u_{ws} \\ k=1,\dots,K}} \frac{1}{\sigma(u_{wk}^T v_c)} (1-\sigma(u_{wk}^T v_c)) \cdot \sigma(-u_{wk}^T v_c) \cdot (-v_c)$

$\frac{\partial J_{neg}}{\partial u_{ws}} = \sum_{\substack{u_{wk}=u_{ws} \\ k=1,\dots,K}} (1-\sigma(-u_{wk}^T v_c) v_c$

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

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

$\frac{\partial J_{skip\text{-}gram}}{\partial u_w} = \sum_{\substack{-m\leq j\leq m \\ j=0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial u_w} = 0 \quad (w \neq c)$


enjoyable
rain
annoying
brilliant
sweet
tea
snow
cool
bad
coffee
woman
female
man
worth
hail
amazing
wonderful
male
boring
great
queen
dumb
well
king
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.1
0.0
0.1
0.2
0.3

The illustration demonstrates that certain words are grouped together, but they may not always be located in the same location.

For example, the words "amazing," "wonderful," "great," and "boring" are grouped together to represent e motions, as are "cool" and "brilliant," and "enjoyable" and "annoying."

However, some of the groupings are more successful than others. For example, the words "female" and " woman," as well as "tea" and "coffee," are more closely clustered together.