Baktash Ansari

a) Based on explanations above y is one-hot vector which has a 1 for the true outside word o, and 0 every-where else, so all terms will be zero except for $w = o$

so we have: $-\sum_{w \in V} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$

b) Dimensions of u and v is ~~vector size~~ number of words for shape(0) and one for shape(1) (num, 1) so they are same in Dimension and we have:

$$\frac{\partial J}{\partial v_c} = \frac{-\partial \log P(o|c)}{\partial v_c} = -\partial \log \left(\frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} \right)$$

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

$$= \frac{-\partial (u_o^T v_c)}{\partial v_c} + \frac{1}{\sum_{w \in V} e^{u_w^T v_c}} \times \frac{\partial \left(\sum_{w \in V} e^{u_w^T v_c} \right)}{\partial v_c}$$

$$= -u_o^T + \sum_{w \in V} \frac{e^{u_w^T v_c}}{\sum_{w \in V} e^{u_w^T v_c}} \times u_w^T$$

$$= -u_o^T + \sum_{w \in V} P(O=w \mid C=c) \times u_w^T = -u_o^T + \hat{y} \times v^T$$

$$= \underline{v^T(\hat{y} - y)}$$

ii| when is the gradient you computed equal to zero?

based on answer in last part $v^T(\hat{y} - y) = 0$

so we have : $\hat{y} = y \Rightarrow$ we can understand that gradient will be equal to zero whenever our estimated word equal to real word.

iii| when we subtract gradients from v_c we are creating Gradient Descent formula :

$$\theta^{new} = \theta^{old} - \alpha \nabla_\theta J(\theta)$$

that we use it when we want to minimize the value of loss function which lead to predict true word with higher probability.

iv) Regarding to hint, in our data if we have some words that similar word vectors but not identical (like synonym words) by L2 normalization these vectors consider equall and can make noise in our calculations. or for example if sentiment of a phrase is important for us and we have two phrases which have different emotions but similar word vectors like (very good, kind of good), L2 may remove this informations. on the other hand in tasks that high similarity is not so important, we can use it.

c) when $w=o$ &
$$\frac{\partial(J(v_c,o,U))}{\partial v} = \frac{\partial(-\log P(O=o|C=c))}{\partial v}$$

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

$$= -\frac{\partial u_o^T v_c}{\partial v} + \sum_{w\in V} \frac{e^{u_w^T v_c}}{\sum_{w\in V} e^{u_w^T v_c}} x v_c^T$$

$$= -\frac{\partial u_o^T v_c}{\partial v} + \sum_{w\in V} P(O=w|C=c)\, x\, v_c^T$$

$$= -\frac{\partial v_y^T v_c}{\partial v} + \hat{y} v_c^T = -y^T v_c + \hat{y} v_c^T = v_c^T(\hat{y}-y)$$

if $w=0 \Longrightarrow y=1 \Longrightarrow$ Result $= \overline{V_c^T(\hat{y}-1) = V_c^T\hat{y} - V_c^T}$

O.W $\Longrightarrow y=0 \Longrightarrow$ Result $= \overline{V_c^T\hat{y}}$

d) $\frac{\partial J}{\partial u} = \left[\frac{\partial J(V_c,O,U)}{\partial U_1}, \frac{\partial J(V_c,O,U)}{\partial U_2}, \dots, \frac{\partial J(V_c,O,U)}{\partial U_{|Vocab|}}\right]$

e) Based on Leaky ReLU:

$$f(x) = \max(\alpha x, x) \quad (\alpha < 1) = \begin{cases} \alpha x & x<0 \\ x & x>0 \end{cases}$$

so we have: $\frac{\partial f(x)}{\partial x}(x<0) = \alpha$

$\frac{\partial f(x)}{\partial x}(x>0) = 1$

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

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

d)

i) $$\frac{\partial(J(v_c, o, U))}{\partial v_c} = \frac{\partial(-\log(\sigma(u_o^T v_c)))}{\partial v_c} - \sum_{s=1}^{K} \frac{\partial(\log(\sigma(-u_{w_s}^T v_c)))}{\partial v_c} = -\frac{1}{\sigma(u_o^T v_c)} \times \sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))$$

$$\times u_o^T + \sum_{s=1}^{K} \frac{1}{\sigma(-u_{w_s}^T v_c)} \times \sigma(-u_{w_s}^T v_c)(1-\sigma(-u_{w_s}^T v_c)) \times u_{w_s}^T$$

$$= (\sigma(u_o^T v_c) - 1) u_o^T + \sum_{s=1}^{K} (1-\sigma(-u_{w_s}^T v_c)) u_{w_s}^T$$

$$= u_o^T \sigma(u_o^T v_c) - u_o^T + \sum_{s=1}^{K} (u_{w_s}^T - u_{w_s}^T \sigma(-u_{w_s}^T v_c))$$

$$\frac{\partial(J(v_c, o, U))}{\partial u_o} = \frac{\partial(-\log(\sigma(u_o^T v_c)))}{\partial u_o} - \sum_{s=1}^{K} \frac{\partial(\log(\sigma(-u_{w_s}^T v_c)))}{\partial u_o}$$

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

$$= (\sigma'(u_o^T v_c) - 1) v_c^T$$

$$\frac{\partial(J(v_c, o, U))}{\partial u_{w_s}} = \frac{\partial(-\log(\sigma(u_o^T v_c)))}{\partial u_{w_s}} - \sum_{s=1}^{K} \frac{\partial(\log(\sigma(-u_{w_s}^T v_c)))}{\partial u_{w_s}} = 0 + \sum_{s=1}^{K} \frac{1}{\sigma(-u_{w_s}^T v_c)} \times$$

$$\sigma(-u_{w_s}^T v_c)(1-\sigma(-u_{w_s}^T v_c)) \times v_c^T = (1-\sigma(-u_{w_s}^T v_c)) v_c^T$$

NAHAL

ii) Based on calculations of last part we use computions of first derivative in others so we reuse the $\frac{\partial(J_{neg\text{-}sample})}{\partial v_c}$, for representation we have:

$$\frac{\partial(J_{neg\text{-}sample})}{\partial v_c} = -u_{o\{w_1,\dots,w_K\}}\left(\sigma\left(u_{o\{w_1,\dots,w_K\}}^{T} v_c - 1\right)\right)$$

iii) Because it only compute the gradients for small number of samples but naive-Softmax loss needs to compute gradients for all possible samples.

h) Based on previous computations we have:

$$\frac{\partial J}{\partial w_s} = v_c^{T}\left(1-\sigma\left(-u_{w_s}^{T} v_c\right)\right)$$

but now we know that for word in vocab, some w_s are equal to w_s so we should compute the derivative multiple times, so we have:

$$\frac{\partial J}{\partial w_s} = \sum_{\substack{w \in V \\ w = w_s}}^{K} v_c^{T}\left(1-\sigma\left(-u_{w_s}^{T} v_c\right)\right)$$

i)

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

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

iii) $\frac{\partial(J_{skip\text{-}gram})}{\partial v_w} = 0$