ⓐ Because y is one-hot vector $\rightsquigarrow y_w = \begin{cases} 0 & \text{o.w} \\ 1 & w=o \end{cases}$

So we can remove all $y_w = 0$.

ⓑ i) $\frac{\partial J(v_c, o, U)}{\partial v_c} = -\frac{\partial}{\partial v_c} \log p(O=o \mid C=c)$

$= -\frac{\partial}{\partial v_c} \log \frac{\exp(u_o^T v_c)}{\sum_{w\in vocab} \exp(u_w^T v_c)} = -\frac{\partial}{\partial v_c} \log \exp(u_o^T v_c)$

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

$\frac{1}{\sum_{w\in vocab} \exp(u_w^T v_c)} \frac{\partial}{\partial v_c} \sum \exp(u_w^T v_c) = -u_o +$

$\frac{1}{\sum_{w\in vocab} \exp(u_w^T v_c)} \sum_{w\in vocab} \frac{\partial}{\partial v_c} \exp(u_w^T v_c) = -u_o +$

$+ \frac{1}{\sum_{w\in vocab} \exp(u_w^T v_c)} \sum_{w\in vocab} \exp(u_w^T v_c) \frac{\partial}{\partial v_c} u_w^T v_c$

$= -u_o + \frac{\sum_{w\in vocab} \exp(u_w^T v_c)\, u_w}{\sum_{w\in vocab} \exp(u_w^T v_c)} = -Uy + \sum_{w\in vocab} \frac{\exp(u_w^T v_c)}{\exp(u_k^T v_c)}$

$\times u_w = -Uy + \sum_{w\in vocab} \hat{y}_w u_w = -Uy + U\hat{y} = U(\hat{y} - y)$

$v_c, u_w \in \mathbb{R}^N \quad U \in \mathbb{R}^{N\times V} \quad y, \hat{y} \in \mathbb{R}^{V\times 1} \quad \frac{\partial J}{\partial v_c} \in \mathbb{R}^N$

Scanned with CamScanner

Subject
Date

ii) $U(y-\hat{y}) = 0 \longrightarrow \underbrace{Uy}_{u_o} = U\hat{y}$

The gradient is equal to zero when predicted word is equal to outside word.

iii) We subtract this difference from v_c to get closer to optimum point. It means that we want to have minimum difference between y and $\hat{y}$.

iv) In some cases, L2 normalization might take away useful information that could be relevant for the task. for example, in sentiment analysis where the goal is to classify phrases as positive or negative, it may not be good, because L2 scales the embeddings to have a unit norm & remove information about the original magnitude. Well in tasks such as text similarity or clustering it can be helpful in reducing the influence of outliers.

Subject

Date

(c) $-\frac{\partial}{\partial u_w} \log P(O=o \mid C=c) = -\frac{\partial}{\partial u_w} \log \frac{\exp(u_o^T v_c)}{\sum_{w \in vocab} \exp(u_w^T v_c)}$

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

$= -\frac{\partial}{\partial u_w} u_o v_c + \frac{1}{\sum_w^V \exp(u_w v_c)} \cdot \frac{\partial}{\partial u_w} \sum_w^V \exp(u_w^T v_c)$

$= -\frac{\partial}{\partial u_w} u_o v_c + \frac{1}{\sum_w^V \exp(u_w v_c)} \cdot \frac{\partial}{\partial u_w} \exp(u_w v_c)$

$= -\frac{\partial}{\partial u_w} u_o v_c + \frac{\exp(u_w v_c)}{\sum_w^V \exp(u_w v_c)} \cdot v_c = -y_w v_c + \hat{y}_w v_c$

$= (\hat{y}_w - y_w) v_c^T \qquad y_w = \begin{cases} 1 & w=o \\ 0 & o.w \end{cases}$

(d) $\frac{\partial J}{\partial U} = \left[\frac{\partial J}{\partial u_1}, \frac{\partial J}{\partial u_2}, \dots, \frac{\partial J}{\partial u_{|vocab|}} \right]$

(e)

$f(x) = \begin{cases} x & x \geq 0 \\ \alpha x & x < 0 \end{cases} \qquad f'(x) = \begin{cases} 1 & x > 0 \\ \alpha & x < 0 \end{cases}$

(f) $\frac{d}{dx} \frac{1}{1+e^{-x}} = \frac{(-1)(-1)e^{-x}}{(1+e^{-x})^2} = \frac{1+e^{-x}}{(1+e^{-x})^2} - \frac{1}{(1+e^{-x})^2}$

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

Subject

Date

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

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

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

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

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

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

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

* $\frac{\partial J}{\partial u_k} = -\frac{\partial}{\partial u_k} \log(\sigma(u_o^T v_c)) - \frac{\partial}{\partial u_k} \sum_{x=1}^{K} \log(\sigma(-u_x^T v_c))$

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

ii)

$U_{o,\{w_1,\dots,w_K\}} = [u_o, -u_{w_1}, \dots, -u_{w_K}]$

$U_o^T v_c = \begin{bmatrix} u_o^T v_c \\ \vdots \\ -u_{w_K}^T v_c \end{bmatrix} \quad 1 - \sigma(U_o^T \cdot v_c) = \begin{bmatrix} 1 - \sigma(u_o^T v_c) \\ \vdots \\ 1 - \sigma(u_{w_K}^T v_c) \end{bmatrix}$

Subject

Date

iii) Rather than computing the probabilities for all possible output words, negative sampling only considers a small number of negative samples per positive sample, greatly reducing the number of computations.

(h) $\frac{\partial J}{\partial u_k} = -\frac{\partial}{\partial u_k} \log(\sigma(u_o^T v_c)) - \frac{\partial}{\partial u_k} \sum_{x=1}^{K} \log(\sigma(-u_x^T v_c))$

$= -\frac{\partial}{\partial u_k} \sum_{\substack{u_x = u_k \\ x=1,\dots,k}} \log(\sigma(-u_x^T v_c)) - \frac{\partial}{\partial u_k} \sum_{\substack{u_x \neq u_k \\ x=1,\dots,k}} \log(\sigma(-u_x^T v_c))$

$= -\sum_{\substack{u_x = u_k \\ x=1,\dots,k}} \frac{\partial}{\partial u_k} \log(\sigma(-u_k^T v_c)) = -\left(\sum_{x=1}^{K} \mathbb{I}\{u_x = u_k\}\right)$

$\frac{\partial}{\partial u_k} \log(\sigma(-u_k^T v_c)) = \left(\sum_{x=1}^{K} \mathbb{I}\{u_x = u_k\}\right)(1-\sigma(-u_k^T v_c)) v_c$

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

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

iii) $$\frac{\partial J_{\text{skip-gram}}(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial v_w} = 0 \quad w \neq c$$