غزل زمانی نژاد
401722244

a) as it is mentioned before, y is a one-hot vector with a 1 for true outside word & 0 elsewhere so:

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\left(\underbrace{y_{w_1}}_{0} \log(\hat{y}_{w_1}) + \underbrace{y_{w_2}}_{0} \log(\hat{y}_{w_2}) + \dots + \underbrace{y_o}_{1} \log(\hat{y}_o) + \dots + \underbrace{y_{w_v}}_{0} \log(\hat{y}_{w_v})\right) =$$

$$= -y_o \log(\hat{y}_o) = -\log(\hat{y}_o)$$

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

$$= -\underbrace{\log \exp(u_o^T v_c)}_{A} + \underbrace{\log \sum_{w \in vocab} \exp(u_w^T v_c)}_{B}$$

$$\partial J / \partial v_c = -\frac{\partial A / \partial v_c}{A} + \frac{\partial B / \partial v_c}{B} = \frac{-u_o \exp(u_o^T v_c)}{\exp(u_o^T v_c)} + \frac{\sum_{w \in V} u_w^T \exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} =$$

$$= -u_o + \sum_{w \in V} u_w^T \underbrace{\frac{\exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}}_{\to P(O=o \mid C=c)} = -u_o + \sum_{w \in vocab} u_w^T \hat{y}_w = U(\hat{y} - y)$$

c) Case 1 - $w=o$: $$\partial J / \partial u_w = -\frac{\partial A / \partial u_w}{A} + \frac{\partial B / \partial u_w}{B} = \frac{-v_c \exp(u_o^T v_c)}{\exp(u_o^T v_c)} +$$

$$\frac{v_c \exp(u_w^T v_c)}{\sum_{w \in vocab} \exp(u_w^T v_c)} = -v_c + v_c \hat{y}_w$$

Case 2 - $w \neq o$: $$\partial J / \partial u_w = \frac{\overbrace{\partial / \partial u_w \exp(u_o^T v_c)}^{\to 0}}{A} + v_c \hat{y}_w = v_c \hat{y}_w$$

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

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

$$= \sigma(x) \times (-)\, \sigma(x)\sigma(x) = \sigma(x)(1-\sigma(x))$$

f) $f(u) = \log(u)$, $g(u) = \sigma(u)$ $\quad (f(g(u)))' = f'(g(u)) \cdot g'(u)$

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

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

$$\partial J / \partial u_o = -\frac{v_c\, \sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))}{\sigma(u_o^T v_c)} - \underbrace{0}_{o \notin K} = -v_c(1-\sigma(u_o^T v_c))$$

$$\partial J / \partial u_k = 0 - (-v_c)(1-\sigma(-u_k^T v_c)) = v_c(1-\sigma(-u_k^T v_c))$$

g) $$J = -\log \sigma(u_o^T v_c) - \sum_{u_i = u_k}^{Count} \log \sigma(-u_i^T v_c) - \sum_{u_j \neq u_k} \log \sigma(-u_j^T v_c)$$

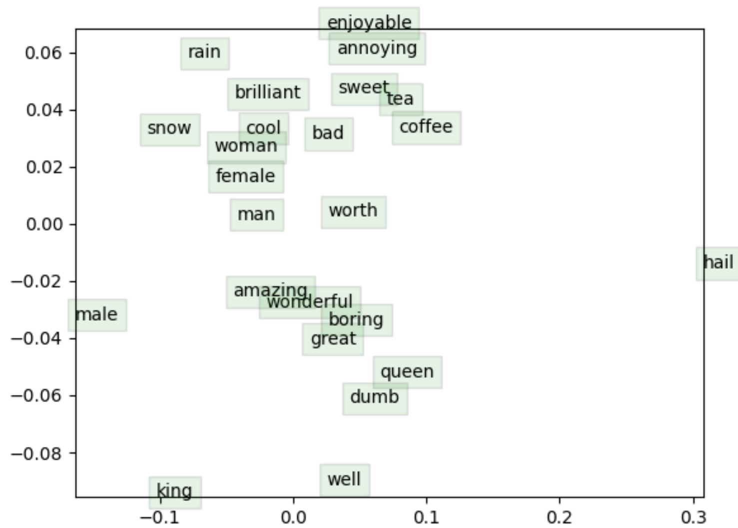
$$\partial J / \partial u_k = 0 - Count \frac{-v_c\, \sigma(-u_k v_c)(1-\sigma(-u_k^T v_c))}{\sigma(-u_k^T v_c)} - 0 = Count \times v_c(1-\sigma(-u_k^T v_c))$$

h) $$\partial J / \partial U = \sum_{-m \leq j \leq m} \partial / \partial U \; J(v_c, w_{t+j}, U)$$

$$\partial J / \partial v_c = \sum_{-m \leq j \leq m} \partial / \partial v_c \; J(v_c, w_{t+j}, U)$$

$$\partial J / \partial v_{w (w \neq c)} = 0$$

Coding part:



As it is seen, related words are almost clustered together: rain, snow/ tea, coffee/ amazing, wonderful, boring, great. Moreover, the semantic pattern is almost seen in this plot: male as king, female as queen. Some of the embeddings are not as good as expected (for example male should be clustered with woman, female, man); it is related to the corpus that we have trained on.