< Assignment 01 > 20225472 박상준 (Sangjune Park)

#1 - (a). Because only o is the position that the value is 1, and 0 everywhere else.

#1 - (b).

Substituting (1) into (2), then

$$J = \log \sum_w \left(\exp(u_w^T v_c)\right) - u_o^T v_c$$

Here, $$\frac{\partial J}{\partial v_c} = -u_o + \frac{\sum_w \overbrace{\exp(u_w^T v_c)}^{\hat{y}_w} \cdot u_w}{\sum_w \exp(u_w^T v_c)} = -u_o + \sum_w \hat{y}_w \cdot u_w$$

$$U = \begin{bmatrix} | & | & & | & \\ u_1 & u_2 & \cdots & u_w & \cdots \\ | & | & & | & \end{bmatrix}$$ and y is one-hot vector, so Uy indicates u_o

Also, $U\hat{y}$ indicates $\sum_w \hat{y}_w \cdot u_w$. Therefore, $\underline{\frac{\partial J}{\partial v_c} = U(\hat{y} - y)}$ //

(1) The gradient will be zero if $\hat{y} \approx y$.

(2) By subtracting the gradient from v_c, v_c will be adjusted to the direction that reduce the difference between predicted value and true value. In other words, v_c will be closer to the outside word vectors in its context.

#1 - (c).

i) $w = o$ $(u_w = u_o)$

$$\frac{\partial J}{\partial u_w} = -v_c + \frac{\overbrace{\exp(u_w^T v_c)}^{\hat{y}_w = \hat{y}_o} \cdot v_c}{\sum_w \exp(u_w^T v_c)} = \underline{(\hat{y}_o - 1) v_c} //$$

ii) $w \neq o$

$$\frac{\partial J}{\partial u_w} = \frac{\overbrace{\exp(u_w^T v_c)}^{\hat{y}_w} \cdot v_c}{\sum_w \exp(u_w^T v_c)} = \underline{\hat{y}_w v_c} //$$

#1 - (d). With the answer of 1-(c), and assuming $V = |\text{Vocab}|$,

$$\frac{\partial J}{\partial u_1} = \begin{cases} (\hat{y}_1 - 1) v_c & \text{if } w = 1 = o \\ \hat{y}_1 v_c & \text{if } w = 1 \neq o \end{cases} \quad \frac{\partial J}{\partial u_2} = \begin{cases} (\hat{y}_2 - 1) v_c & \text{if } w = 2 = o \\ \hat{y}_2 v_c & \text{if } w = 2 \neq o \end{cases} \quad \cdots \quad \frac{\partial J}{\partial u_V} = \begin{cases} (\hat{y}_V - 1) v_c, & \text{if } w = V = o \\ \hat{y}_V v_c, & \text{if } w = V \neq o \end{cases}$$

1-(e).

$$\frac{df}{dx} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$$

1-(f).

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

1-(g).

$$J = -\log\left(\sigma(u_o^T v_c)\right) - \sum_{s=1}^{K} \log\left(\sigma(-u_{w_s}^T v_c)\right)$$

(i). For v_c,

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

$$= -\left(1-\sigma(u_o^T v_c)\right) u_o + \sum_s \left(1-\sigma(-u_{w_s}^T v_c)\right) u_{w_s}$$

For u_o,

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

For u_{w_s},

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

(ii)

$$U_{o,\{w_1,\cdots,w_K\}}^T \cdot v_c = \left[u_o^T v_c,\ -u_{w_1}^T v_c,\ \cdots,\ -u_{w_K}^T v_c\right]$$

Therefore, we can reuse $\sigma(U_{o,\{w_1,\cdots,w_K\}}^T \cdot v_c) - \mathbb{1}$.

(iii) We need lesser vectors than the naive-softmax loss.

1-(h).

$$J = -\log\left(\sigma(u_o^T v_c)\right) - \sum_{w_j = w_s} \log\left(\sigma(-u_{w_j}^T v_c)\right) - \sum_{w_j \neq w_s} \log\left(\sigma(-u_{w_j}^T v_c)\right)$$

Assuming there are m w_s words, $\frac{\partial J}{\partial w_s} = m\left(1-\sigma(-u_{w_s}^T v_c)\right) v_c$

(i)

$$\frac{\partial J(v_c, w_{t-m}, \dots, w_{t+m}, U)}{\partial U} = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

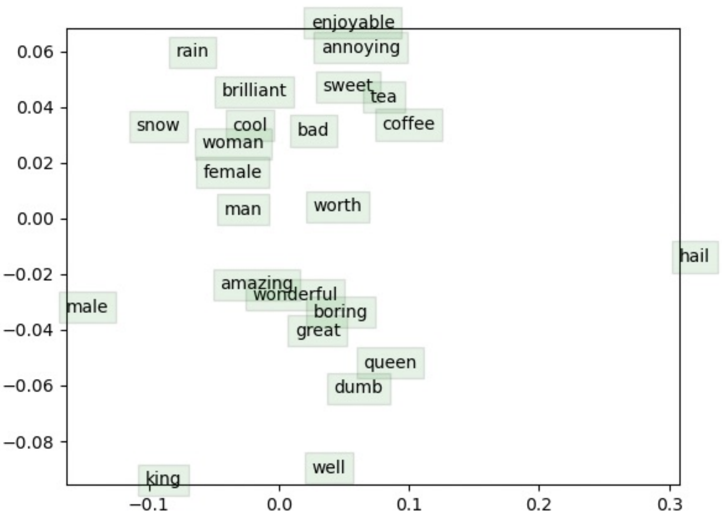
(ii)

$$\frac{\partial J(v_c, w_{t-m}, \dots, w_{t+m}, U)}{\partial v_c} = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

(iii)

$$\frac{\partial J(v_c, w_{t-m}, \dots, w_{t+m}, U)}{\partial v_w} = 0$$

#2-(C).



There are many well-clustered words such as (amazing, wonderful, boring, great). However, there are also not well-clustered words like "hail", which should have clustered with (rain, snow). Also, in evaluation, antonyms are considered as similar, and it is shown in the cluster "great (pos) ↔ boring (neg)"