CS224-N Assignment-2

1. Understanding word2vec

(a) To prove:

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$$

y is an one-hot vector with 1 in context vector position and 0 in all other positions.

$$y_w = 1 \quad \text{if } w \text{ is context word}$$
$$= 0 \quad \text{otherwise.}$$

$$\therefore -\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -[y_1 \log(\hat{y}_1) + \dots + y_o \log(\hat{y}_o) + \dots y_v \log(\hat{y}_v)]$$
$$= -[0 + \dots 1 \cdot \log(\hat{y}_o) + \dots 0]$$
$$= -\log \hat{y}_o$$

(b) To find: $\frac{\partial J_{ns}(v_c, o, U)}{\partial v_c} = ?$

$$J_{ns}(v_c, o, U) = -\log P(O=o|C=c)$$
$$= -\log \frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}$$
$$= -[\log \exp(u_o^T v_c) - \log \sum_{w \in Vocab} \exp(u_w^T v_c)]$$
$$= -[u_o^T v_c - \log \sum_{w \in Vocab} \exp(u_w^T v_c)]$$

$$\frac{\partial J(v_c, o, U)}{\partial v_c} = \frac{\partial}{\partial v_c}\left[\log \sum_{w \in Vocab} \exp(u_w^T v_c) - u_o^T v_c\right] = \underbrace{\frac{\partial}{\partial v_c}[\log \sum_w \exp(u_w^T v_c)]}_{1} - \underbrace{\frac{\partial}{\partial v_c}[u_o^T v_c]}_{2}$$

(2) $\rightarrow \frac{\partial}{\partial v_c}[u_o^T v_c] = u_o.$

(1) $\rightarrow \frac{\partial}{\partial v_c}[\log \sum_w \exp(u_w^T v_c)] = \frac{1}{\sum_w \exp(u_w^T v_c)} \sum_{x \in Vocab} \frac{\partial}{\partial v_c} \exp(u_x^T v_c).$

$$= \frac{1}{\sum_w \exp(u_w^T v_c)} \sum_{x \in Vocab} \exp(u_x^T v_c) \frac{\partial}{\partial v_c}(u_x^T v_c)$$

$$= \frac{1}{\sum_w \exp(u_w^T v_c)} \sum_{x \in Vocab} \exp(u_x^T v_c)\, u_x$$

$$= \frac{\sum_{x \in Vocab} \exp(U_x^T V_c)\, U_x}{\sum_{w \in Vocab} \exp(U_w^T V_c)} = \sum_{x \in Vocab} P(x|c)\, U_x = \sum_{x \in Vocab} \hat{y}_x \cdot U_x$$

$$\frac{\partial J_{ns}(V_c, 0, U)}{\partial V_c} = ① - ② = \cancel{\phi} \cancel{\sum_{x \in Vocab} \hat{y}_x U_x}$$

$$= \sum_{x \in Vocab} \hat{y}_x \cdot U_x - U_o = \sum_{x \in Vocab} \hat{y}_x \cdot U_x - \sum_{w \in Vocab} y_w \cdot U_w$$

ⓒ $\frac{\partial J_{ns}(V_c, o, U)}{\partial U_i} = ?$ when $i = o$ and $i \neq o$.

when $i = o$

$$\frac{\partial J_{ns}(V_c, 0, U)}{\partial U_o} = \frac{\partial}{\partial U_o}\left[\log \sum_{w \in Vocab} \exp(U_w^T V_c) - U_o^T V_c\right]$$

$$= \underbrace{\frac{\partial}{\partial U_o}\left[\log \sum_{w \in Vocab} \exp(U_w^T V_c)\right]}_{①} - \underbrace{\frac{\partial}{\partial U_o}\left[U_o^T V_c\right]}_{②}$$

$$② \rightarrow \frac{\partial}{\partial U_o}\left[U_o^T V_c\right] = V_c$$

$$① \rightarrow \frac{\partial}{\partial U_o}\left[\log \sum_{w \in Vocab} \exp(U_w^T V_c)\right] = \frac{1}{\sum_w \exp(U_w^T V_c)} \cdot \sum_{x \in Vocab} \frac{\partial}{\partial U_o} \exp(U_x^T V_c)$$

$$= \frac{1}{\sum_w \exp(U_w^T V_c)} \cdot \left[\cancel{\frac{\partial}{\partial U_o} \exp(U_1^T V_c)}^{0} + \dots \frac{\partial}{\partial U_o} \exp(U_o^T V_c) + \dots \cancel{\frac{\partial}{\partial U_o} \exp(U_v^T V_c)}^{0}\right]$$

$$= \frac{1}{\sum_w \exp(U_w^T V_c)} \cdot \exp(U_o^T V_c) \cdot \frac{\partial}{\partial U_o}(U_o^T V_c).$$

$$= \frac{\exp(U_o^T V_c)}{\sum_w \exp(U_w^T V_c)} \cdot V_c = P(o|c) \cdot V_c$$

$$\frac{\partial}{\partial U_o} J_{ns}(V_c, 0, U) = ① - ② = P(o|c) \cdot V_c - V_c = \sum_{x \in Vocab} y_x \cdot \hat{y}_x \cdot V_c - V_c$$

When $i \neq 0$.

$$\frac{\partial J_{ns}(v_c, 0, U)}{\partial U_i} = \frac{\partial}{\partial U_i}\left[\log \sum_{w \in Vocab} \exp(U_w^T V_c) - U_0^T V_c\right]$$

$$= \frac{\partial}{\partial U_i}\left[\log \sum_{w \in Vocab} \exp(U_w^T V_c)\right] - \frac{\partial}{\partial U_i}(U_0^T V_c)$$

① ②.

$$② \rightarrow \frac{\partial}{\partial U_i}(U_0^T V_c) = 0$$

$$① \rightarrow \frac{\partial}{\partial U_i}\left[\log \sum_{w \in Vocab} \exp(U_w^T V_c)\right] = \frac{1}{\sum_{w \in Vocab} \exp(U_w^T V_c)} \cdot \sum_{x \in Vocab} \frac{\partial}{\partial U_i} \exp(U_x^T V_c)$$

$$= \frac{1}{\sum_{w \in Vocab} \exp(U_w^T V_c)} \cdot \left[\frac{\partial}{\partial U_i} \cancel{\exp(U_1^T V_c)}^{0} + \dots \frac{\partial}{\partial U_i} \exp(U_i^T V_c) + \dots \frac{\partial}{\partial U_i} \cancel{\exp(U_v^T V_c)}^{0}\right]$$

$$= \frac{1}{\sum_{w \in Vocab} \exp(U_w^T V_c)}\left[\frac{\partial}{\partial U_i} \exp(U_i^T V_c)\right] = \frac{1}{\sum_{w} \exp(U_w^T V_c)} \cdot \exp(U_i^T V_c) \cdot \frac{\partial}{\partial U_i}(U_i^T V_c)$$

$$= \frac{\exp(U_i^T V_c)}{\sum_{w \in Vocab} \exp(U_w^T V_c)} \cdot V_c = P(i|c) \cdot V_c$$

$$\frac{\partial J_{ns}(V_c, 0, U)}{\partial U_i} = ① - ② = P(i|c) \cdot V_c - 0$$

$$= \hat{y}_i \cdot V_c$$

(d) $\sigma(x) = \dfrac{1}{1+e^{-x}}$

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

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

$$\sigma'(x) = \sigma(x)[1 - \sigma(x)]$$

c) $J_{neg\text{-}sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$

$$\frac{\partial J_{ns}(v_c, o, U)}{\partial v_c} = \frac{\partial}{\partial v_c}\left[-\log(\sigma(u_o^T v_c))\right]_{①} - \frac{\partial}{\partial v_c}\left[\sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))\right]_{②}$$

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

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

$$② \rightarrow \quad \sum_{k=1}^{K} \frac{\partial}{\partial v_c} \log(\sigma(-u_k^T v_c)) = \sum_{k=1}^{K} \frac{1}{\sigma(-u_k^T v_c)} \sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c)) \cdot \frac{\partial}{\partial v_c}(-u_k^T v_c)$$

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

$$\frac{\partial J_{ns}(v_c, o, U)}{\partial v_c} = ① - ②$$

$$\frac{\partial J_{ns}}{\partial v_c} = -\frac{1}{\cancel{\sigma(u_o^T v_c)}} \cancel{\sigma(u_o^T v_c)}(1-\sigma(u_o^T v_c))\, u_o + \sum_{k=1}^{K} \frac{1}{\cancel{\sigma(-u_k^T v_c)}} \cdot \cancel{\sigma(-u_k^T v_c)} \cdot (1-\sigma(-u_k^T v_c)) \cdot u_k$$

Similarly,

$$\frac{\partial J_{ns}(v_c, o, U)}{\partial u_o} = -\frac{1}{\cancel{\sigma(u_o^T v_c)}} \cancel{\sigma(u_o^T v_c)}(1-\sigma(u_o^T v_c)) \cdot v_c + \frac{1}{\cancel{\sigma(-u_o^T v_c)}} \cdot \cancel{\sigma(-u_o^T v_c)} \cdot (1-\sigma(-u_o^T v_c)) \cdot v_c$$

$$\frac{\partial J_{ns}(v_c, o, U)}{\partial u_k} = 0 + \frac{1}{\cancel{\sigma(-u_k^T v_c)}} \cancel{\sigma(-u_k^T v_c)} \cdot (1-\sigma(-u_k^T v_c)) \cdot v_c$$

$$\frac{\partial J_{ns}}{\partial v_c} = -(1-\sigma(u_o^T v_c))\, u_o + \sum_{k=1}^{K} (1-\sigma(-u_k^T v_c)) \cdot u_k$$

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

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

This loss fn. is more efficient than naive softmax because in naive softmax we have compute $\left(\sum_{w \in vocab} u_w^T v_c\right)$ everytime which is computationally expensive, whereas in negative sampling only the sampled 'k' vectors are used thus resulting in faster computation.

f) $J_{skip\text{-}gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U) = \sum_{\substack{-m \le j \le m \\ j \ne 0}} J(v_c, w_{t+j}, U)$

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

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

iii) $$\frac{\partial J_{SG}(v_c, w_{t-m}, \ldots w_{t+m}, U)}{\partial v_w} \text{ when } w \ne c = 0$$ (as the loss function depends only on v_c and $u_1 \ldots u_k$)