— 1821098008 王 [illegible]

Assignment 4

I word2vec

(a) derive the gradients with respect to V_c.

$$\log P(o|c) = \log \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)} = \log \exp(u_o^T v_c) - \log \sum_{w=1}^{V} \exp(u_w^T v_c)$$

$$\frac{\partial}{\partial v_c} \log \exp(u_o^T v_c) = \frac{\partial}{\partial v_c} u_o^T v_c = u_o$$

$$\frac{\partial}{\partial v_c} \log \sum_{w=1}^{V} \exp(u_w^T v_c) = \frac{1}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \frac{\partial \sum_{w=1}^{V} \exp(u_w^T v_c)}{\partial v_c}$$

$$= \frac{1}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \cdot \sum_{w=1}^{V} \exp(u_w^T v_c) \cdot u_w$$

$$\frac{\partial \log P(o|c)}{\partial v_c} = u_o - \frac{\sum_{w=1}^{V} \exp(u_w^T v_c) \cdot u_w}{\sum_{w=1}^{V} \exp(u_w^T v_c)}$$

$$= u_o - \sum_{t=1}^{V} \frac{\exp(u_t^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \cdot u_t$$

$$= u_o - \sum_{t=1}^{V} P(t|c) u_t$$

$$\frac{\partial J_{softmax-CE}(o, v_c, U)}{\partial v_c} = -\frac{\partial \log P(o|c)}{\partial v_c}$$

$$= -u_o + \sum_{t=1}^{V} P(t|c) u_t$$

(b) 当 $w \neq o$ 时,

$$\frac{\partial u_o^T v_c}{\partial u_w} = 0$$

$$\frac{\partial \log \sum_{w=1}^{V} \exp(u_w^T v_c)}{\partial u_w} = \frac{1}{\sum_{t=1}^{V} \exp(u_t^T v_c)} \cdot \frac{\partial \sum_{w=1}^{V} \exp(u_w^T v_c)}{\partial u_w}$$

$$= \frac{1}{\sum_{t=1}^{V} \exp(u_t^T v_c)} \cdot \frac{\partial \exp(u_w^T v_c)}{\partial u_w}$$

$$= \frac{1}{\sum_{t=1}^{V} \exp(u_t^T v_c)} \cdot \exp(u_w^T v_c) \cdot v_c$$

$$\frac{\partial}{\partial u_w} \log(p(o|c)) = 0 - \frac{\exp(u_w^T v_c)}{\sum_{t=1}^{V} \exp(u_t^T v_c)} \cdot v_c$$

$$\frac{\partial}{\partial u_w} J_{softmax\text{-}CE} = -\frac{\partial}{\partial u_w} \log(p(o|c))$$

$$= \frac{\exp(u_w^T v_c)}{\sum_{t=1}^{V} \exp(u_t^T v_c)} v_c \quad (w \neq o)$$

for $w = o$,

$$\frac{\partial u_o^T v_c}{\partial u_o} = v_c$$

$$\frac{\partial}{\partial u_o} J_{softmax\text{-}CE} = -v_c + \frac{\exp(u_o^T v_c)}{\sum_{t=1}^{V} \exp(u_t^T v_c)} \cdot v_c \quad (w = o).$$

(c)

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

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

$$= (1 - \sigma(u_o^T v_c)) u_o$$

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

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

$$\frac{\partial J_{neg\text{-}sample}}{\partial V_c} = -(1-\sigma(U_o^T V_c))U_o + \sum_{k=1}^{K}(1-\sigma(-U_k^T V_c))U_k$$

$$= (\sigma(U_o^T V_c)-1)U_o + \sum_{k=1}^{K}(1-\sigma(-U_k^T V_c))U_k$$

(c.12)

$$\frac{\partial \log(\sigma(U_o^T V_c))}{\partial U_k} = 0$$

$k \neq 0$ 时,

$$\frac{\partial \sum_{k=1}^{K} \log(\sigma(-U_k^T V_c))}{\partial U_k} = \sum_{k=1}^{K} \frac{\partial \log(\sigma(-U_k^T V_c))}{\partial U_k} = \frac{\partial \log(\sigma(-U_k^T V_c))}{\partial U_k}$$

$$= \frac{1}{\sigma(-U_k^T V_c)} \cdot \sigma(-U_k^T V_c)(1-\sigma(-U_k^T V_c))(-V_c)$$

$$= (1-\sigma(-U_k^T V_c)) \cdot (-V_c)$$

$$\frac{\partial J_{neg\text{-}sample}}{\partial U_k} = (1-\sigma(-U_k^T V_c)) \cdot V_c$$

$k=0$ 时, $\dfrac{\partial \log(\sigma(U_o^T V_c))}{\partial U_o} = \dfrac{1}{\sigma(U_o^T V_c)} \cdot \sigma(U_o^T V_c)(1-\sigma(U_o^T V_c)) \cdot V_c$

$$= (1-\sigma(U_o^T V_c)) \cdot V_c.$$

$$\frac{\partial J_{neg\text{-}sample}}{\partial U_o} = (\sigma(U_o^T V_c)-1)V_c + (1-\sigma(-U_k^T V_c)) \cdot V_c.$$

(c.2)

In the softmax-CE loss. the skip-gram neural network updates times as many as the size of word vocabulary. The cost of computing $P(U_o|V_c)$ is proportional with the vocabulary size V.

In the negative sampling loss, we only update certain number weights which is a small percentage of all the possible words.

②

(d) $$\frac{\partial J_{skip\text{-}gram}(Word_{c-m,\cdots,c+m})}{\partial V_c}$$

$$= -\sum_{-m\le j\le m, j\ne 0} \frac{\partial F(W_{c+j}, V_c)}{\partial V_c}$$

$$\frac{\partial J_{skip\text{-}gram}(Word_{c-m,\cdots,c+m})}{\partial U_{W_{c+j}}}$$

$$= \sum_{-m\le j\le m, j\ne 0} \frac{\partial F(W_{c+j}, V_c)}{\partial U_{W_{c+j}}}$$

$\frac{\partial}{\partial V_c} F(W_{c+j}, V_c)$ 和 $\frac{\partial}{\partial U_{W_{c+j}}} F(W_{c+j}, V_c)$ 在 (a)~(c) 中均已经表示出.