Written part Linsen Li

(a) Softmax

Assume $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1}$, then $\text{softmax}(\vec{x}) = \begin{bmatrix} \vdots \\ \frac{e^{x_i}}{\sum_j e^{x_j}} \\ \vdots \end{bmatrix}_{n\times 1}$, that is, $\text{softmax}(\vec{x})_i = \dfrac{e^{x_i}}{\sum_j e^{x_j}}$ is the i-th entry of the vector.

So, $\text{softmax}(\vec{x}+c)_i = \dfrac{e^{x_i+c}}{\sum_j e^{x_j+c}} = \dfrac{e^{x_i}\cdot e^{c}}{\sum_j e^{x_j}\cdot e^{c}} = \dfrac{e^{c}\cdot e^{x_i}}{e^{c}\cdot \sum_j e^{x_j}}$

Since $e^c \neq 0$, then $\text{softmax}(\vec{x}+c)_i = \dfrac{e^{x_i}}{\sum_j e^{x_j}} = \text{softmax}(\vec{x})_i$

Then, $\text{softmax}(\vec{x}) = \text{softmax}(\vec{x}+c)$ for any vector $\vec{x}$ and constant c.

(b) Sigmoid

Set $M = 1+e^{-x}$, then $\sigma(x) = \dfrac{1}{M}$, $\dfrac{\partial \sigma}{\partial M} = -\dfrac{1}{M^2} = -\dfrac{1}{(1+e^{-x})^2}$

then $\dfrac{\partial M}{\partial x} = e^{-x}\cdot(-1) = -e^{-x}$

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

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

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

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

$$\therefore \sigma'(x) = \sigma(x)(1-\sigma(x))$$

(c) Word2vec

(1) For v_c:

Since y is a one-hot vector, that is $y = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ - ith

then when o is the expected word, $\sum_i y_i \log(\hat{y}_i) = 1 \times \log(\hat{y}_o) = \log(\hat{y}_o)$

$$\therefore J_{CE}(o, v_c, U) = -\log(\hat{y}_o) = -\log \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)}$$

$$= -u_o^T v_c + \log \sum_{w=1}^{W} \exp(u_w^T v_c)$$

$$\frac{\partial u_o^T v_c}{\partial v_c} = u_o \text{, which is trivial}$$

Set $M = \sum_{w=1}^{W} \exp(u_w^T v_c)$, $Y = \log M$.

then $\frac{\partial M}{\partial v_c} = \sum_{w=1}^{W} \exp(u_w^T v_c) \cdot u_w$, $\frac{\partial Y}{\partial M} = \frac{1}{M}$

$$\therefore \frac{\partial Y}{\partial v_c} = \frac{\partial M}{\partial v_c} \cdot \frac{\partial Y}{\partial M} = \sum_{w=1}^{W} \exp(u_w^T v_c) \cdot u_w \Big/ \sum_{w=1}^{W} \exp(u_w^T v_c)$$

$$\therefore \frac{\partial J}{\partial v_c} = -u_o + \sum_{x=1}^{W} \frac{\exp(u_x^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot u_x$$

$$= -u_o + \sum_{x=1}^{W} P(x|c) \cdot u_x$$

Since $\hat{y} = \begin{bmatrix} P(1|c) \\ P(2|c) \\ \vdots \\ P(w|c) \end{bmatrix}$, $U = [u_1, u_2, \ldots u_W]$, $u_o = [u_1, u_2 \ldots u_W] \cdot \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}$ o-th.

$$\therefore \sum_{x=1}^{W} P(x|c) \cdot u_x = U \cdot \hat{y}, \quad -u_o = U \cdot y$$

$$\therefore \frac{\partial J_{CE}(o, v_c, U)}{\partial v_c} = U(\hat{y} - y)$$

(2)

For u_w

Since $J_{CE}(o, v_c, U) = -u_o^T v_c + \log \sum_{w=1}^{W} \exp(u_w^T v_c)$

When $w = o$:

$$\frac{\partial J_{CE}(o, v_c, U)}{\partial u_w} = -v_c + \frac{1}{\sum_{x=1}^{W} \exp(u_x^T v_c)} \times \exp(u_w^T v_c) \times v_c$$

$$= -v_c + v_c \cdot \frac{\exp(u_w^T v_c)}{\sum_{x=1}^{W} \exp(u_x^T v_c)}$$

$$= -v_c + v_c \cdot P(w|c)$$

$$= v_c(\hat{y}_w - 1)$$

When $w \neq o$: the only difference is the first term

$$\frac{\partial J_{CE}(o, v_c, U)}{\partial u_w} = 0 + v_c \cdot P(w|c) = v_c \cdot \hat{y}_w$$

$$\therefore \frac{\partial J_{CE}(w, v_c, U)}{\partial u_w} = \begin{cases} (\hat{y}_w - 1) v_c, & w = o \\ \hat{y}_w v_c, & w \neq o \end{cases}$$

(3)

For negative sampling, and $\sigma'(x) = \sigma(x)(1 - \sigma(x))$

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

① For v_c:

$$\frac{\partial J_{neg\text{-}sample}}{\partial v_c} = -\frac{\sigma(u_o^T v_c) \cdot (1 - \sigma(u_o^T v_c))}{\sigma(u_o^T v_c)} \cdot u_o - \sum_{k=1}^{K} \frac{\sigma(-u_k^T v_c)(1 - \sigma(-u_k^T v_c))}{\sigma(-u_k^T v_c)} \cdot (-u_k)$$

Since $\sigma(\cdot) \neq 0$ for all domain.

$$= u_o\left[\sigma(u_o^T v_c) - 1\right] - \sum_{k=1}^{K} (\sigma(-u_k^T v_c) - 1) \cdot u_k$$

② For u_w:

When $w = 0$: since $0 \notin \{1, 2, \dots K\}$, so u_w now is irrelevant. with 2nd term.

$$\therefore \frac{\partial J_{neg\text{-}sample}}{\partial u_w} = -\frac{\sigma(u_o^T v_c)(1 - \sigma(u_o^T v_c))}{\sigma(u_o^T v_c)} \cdot v_c - 0$$

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

When $w \neq 0$: then u_w is irrelevant with 1st term

$$\therefore \frac{\partial J_{neg\text{-}sample}}{\partial u_w} = \frac{\partial}{\partial u_w}\left(-\sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))\right)$$

$$= \frac{\partial}{\partial u_w}\left[-\log(\sigma(-u_w^T v_c))\right] + \frac{\partial}{\partial u_w}\left[-\sum_{k \neq w} \log(\sigma(-u_k^T v_c))\right]$$

$$= -\frac{\sigma(-u_w^T v_c)(1 - \sigma(-u_w^T v_c))}{\sigma(-u_w^T v_c)} \cdot (-v_c) + 0$$

$$= -(\sigma(-u_w^T v_c) - 1) \cdot v_c, \text{ for all } w = 1, 2, \dots K$$

In conclusion:

$$\frac{\partial J_{neg\text{-}sample}}{\partial v_c} = [\sigma(u_o^T v_c) - 1] \cdot u_o - \sum_{k=1}^{K} (\sigma(-u_k^T v_c) - 1) \cdot u_k$$

$$\frac{\partial J_{neg\text{-}sample}}{\partial u_w} = \begin{cases} (\sigma(u_o^T v_c) - 1) \cdot v_c, & \text{when } w = 0 \\ -[\sigma(-u_w^T v_c) - 1] \cdot v_c, & \text{for all } w = 1, 2, \dots K \\ & \text{and } w \neq 0 \end{cases}$$

(4)

Since we have computed

$$\frac{\partial F(w_i, v_c)}{\partial U} \text{ and } \frac{\partial F(w_i, v_c)}{\partial V_c}$$

So, for skip-gram, the gradients for the cost of one context window are

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

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

$$\frac{\partial J_{\text{skip-gram}}(w_{t-m}, \dots w_{t+m})}{\partial V_{w_{t+j}}} = 0, \text{ for all } j \neq c.$$