$$J_{naive\text{-}softmax}(v_c, o, U)$$

$$= -\log P(\text{Outside}=o \mid \text{Center}=c)$$

$$= -\log \frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}$$

$$= -u_o^T v_c + \log \sum_w \exp(u_w^T v_c)$$

$$\frac{\partial J}{\partial v_c} = -u_o^T + \sum_w \frac{\exp(u_w^T v_c)\, u_w}{\sum \exp(u_w^T v_c)} = -u_o + \sum_w P(O=o \mid C=c)\, u_w$$

$$= -u_o + \sum_w \hat{y}_w u_w \leftarrow$$ y는 답에대해서만 1
이외는 0

$$= -u_o + u_w$$

$$= U(\hat{y} - y)$$

b)

---

c)

$J(v_c, o, U) = -\log P(\text{Outside}=o \mid \text{Center}=c)$ 이고,

$$= -\log \frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)}$$

$$= -u_o^T v_c + \log \sum_w \exp(u_w^T v_c)$$

1) w=outside word,

$$\frac{\partial J}{\partial u_w} = -v_c + \frac{\partial}{\partial u_w} \cdot \log \sum_w \exp(u_w^T v_c)$$

$$\frac{(u_w^T v_c)}{\quad} \cdot \frac{(u_w^T v_c)}{\partial u_w}$$

$$= -v_c + \sum_w \exp(u_w^T v_c) \quad \partial \cdots$$

$$= -v_c + P(O|C) \cdot v_c$$

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

ii) $w \neq$ outside word

$$\frac{\partial J}{\partial u_w} = -\frac{\partial}{\partial u_w}(u_o^T v_c) + \frac{\partial}{\partial u_w}\left(\log \sum_w \exp(u_w^T v_c)\right)$$

$$= \frac{(u_w^T v_c)}{\sum_w \exp(u_w^T v_c)} \cdot \frac{\partial}{\partial u_w}(u_w^T v_c) + 0$$

$$= v_c \cdot P(o|c)$$

$$= \hat{y} v_c$$

---

d)

$$\frac{\partial J}{\partial u_1} = \hat{y} v_c$$

$$\frac{\partial J}{\partial u_n} = (\hat{y} - y) v_c \quad (u_w = \text{outsideword})$$

---

e)

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

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

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

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

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

$$\sigma(x) \qquad \sigma(x)$$

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

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

f

$\sigma(\cdot) = \text{sigmoid-function}$

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

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

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

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

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

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

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

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

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

→ negative 이기 때문에

$w \neq o$

$$\therefore \frac{\partial}{\ } (u_o^T v_c)$$

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

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

iii) $$\frac{\partial J}{\partial u_k} = \boxed{-\frac{1}{\sigma(u_o^T v_c)} \cdot \frac{\partial}{\partial u_k}\sigma(u_o^T v_c)}$$

$\rightarrow$ negative sampling, $\{w_1 \cdots w_n\} \in O$ $\therefore 0$

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

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

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

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

$$\frac{\partial J}{\partial u_k} = \underbrace{-\frac{1}{\sigma(u_o^T v_c)} \cdot \frac{\partial}{\partial u_k}\sigma(u_o^T v_c)}_{a} \underbrace{-\sum_{k=1}^{K} \frac{1}{\sigma(-u_k^T v_c)} \cdot \frac{\partial}{\partial v_c}\sigma(-u_k^T v_c)}_{b}$$

u_k = negative sample
$\therefore a = 0$

i) $\forall u = u_k$

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

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

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

$J_{skip\text{-}gram}(v_c, w_{t-m}, \dots w_{t+m}, U)$ f

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

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

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

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

A3

1-(a)

i) $m \leftarrow \beta_1 m + (1-\beta_1)\nabla_\theta J_{minibatch}(\theta)$ 에서, $0<\beta_1<1$ 이므로

각 최적화 step의 size가 계속해서 작아지고, 특정구간의 기울기값이 0 (minimum)인 경우가 발생하지 않는다

ii) v는 gradiant와 비례한다. $\theta \leftarrow \theta - \alpha m/\sqrt{v}$ 이기 때문에

gradiant가 클수록 θ는 작아지게 되고, 반대로 gradiant가 작을수록 θ는 커진다.

이 때문에 gradiant가 작은 위치에서 빠르게 벗어난다.

1-(b)

i) $h_{drop} = \gamma d \odot h$, $E_{p_{drop}}[h_{drop}]_i = h_i$

$\rightarrow E_{p_{drop}}[\gamma d \odot h]_i = h_i$

$d \in \{0, 1\}^{D_h}$

$$\rightarrow d(h_i) = \begin{cases} 0 \rightarrow \text{dropout } O & \rightarrow p_{drop} = 0 \\ h_i \rightarrow \text{dropout } X & \rightarrow p_{drop} = (1 - p_{drop}) \end{cases}$$

$\therefore P(\underset{\text{dropout}}{A}) = p_{drop}$

$$\begin{aligned} E_{p_{drop}}[h_{drop}]_i &= E_{p_{drop}}[\gamma d \odot h] \\ &= \gamma E_{p_{drop}}[d(h_i)]_i \\ &= \gamma \cdot (1 - p_{drop}) \cdot h_i = h_i \\ &\rightarrow \gamma (1 - p_{drop}) = 1 \\ &= \gamma = \frac{1}{1 - p_{drop}} \end{aligned}$$

ii) Dropout은 train data를 사용한 학습 시간 감소 및 train data에 대한 overfitting을 해소할 수 있다. 단, evaluation 상황에서는 모든 뉴런을 사용해 높은 성능을 보인다.

2 - (a)

Root I parsed this sentence correctly

Stack	Buffer	New dependency	Transition
[Root]	[I, parsed, this, sentence correctly]		Initial Configuration
[Root, I]	[parsed, this, sentence correctly]		shift
[Root, I, parsed]	[this, sentence, correctly]		shift
[Root, parsed]	[this, sentence, correctly]	parsed → I	Left-arc
Root, Parsed, this	sentence, correctly		shift
Root, Parsed, this, sentence	correctly		shift
Root, parsed	correctly	sentence → this	Left-Arc
Root, Parsed	correctly	parsed → sentence	Right-Arc
Root, Parsed, correctly			Shift
Root,		Parsed → Correctly	Right-Arc
		Root → Parsed	Right-Arc

→ 2n 번의 연산이 필요