

CS 224n Assignment #2 : word2vec

1 Written: Understanding word2vec (23 points)

<https://jeongukjae.github.io/posts/cs224n-assignments/>

[
 $o = u$ = outside word vector
 $v = c$ = center word vector

• 목적함수 (naive-softmax)

$$J(v_c, o, U) = -\log P(o=o | C=c)$$

$$\bullet P(o=o | C=c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in \text{vocab}} \exp(u_w^T v_c)}$$

(:: skip gram)

• 실제 y : one-hot vector
 \hat{y} : $P(o|C=c)$ (probability)

(a) 공식(2)와 y, \hat{y} 간의 cross-entropy 같다는 것 증명

$$-\sum_w y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$$

: y 는 실제 확률분포, \hat{y} 는 모델이 추정한 확률분포일 때 y 는 context word, o 에 해당하는 element만 1인 one-hot vector이기 때문에 참이된다. (실제값 제외 모두 0이됨)

(b) $J_{\text{naive-softmax}}(v_c, o, U)$, v_c 에 대해 미분하여 y, \hat{y}, U 에 대한 답으로 표현해라

$$i) J = -\log P(o=o | C=c) = -\log \left\{ \frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)} \right\} = -\log(\hat{y}_o)$$

$$P(o=o | C=c) = \frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)} \quad \begin{matrix} \rightarrow \text{center word } (c) \text{ 가 주어졌을 때,} \\ \text{outside } (o) \text{ 가 주어질 확률} \end{matrix}$$

ii) $J(v_c, o, U)$ 에 대해 정리,

$$\begin{aligned} J &= -\log (\exp(u_o^T v_c)) + \log \sum_w \exp(u_w^T v_c) \\ &= -u_o^T v_c + \log \sum_w \exp(u_w^T v_c) \end{aligned} \quad \begin{matrix} \rightarrow \log(e^{u_1^T v_c} + e^{u_2^T v_c} + \dots + e^{u_n^T v_c}) \xrightarrow{n} \\ \frac{1}{e^{u_i^T v_c}} \cdot e^{u_i^T v_c} \cdot u_i^T \end{matrix}$$

$$\begin{aligned} iii) \frac{\partial J}{\partial v_c} &= -u_o + \sum_w \frac{\exp(u_w^T v_c) u_w}{\sum_w \exp(u_w^T v_c)} = -u_o + \sum_w P(o=w | C=c) u_w \\ &= -u_o + \sum_w \hat{y}_w u_w \quad : y \text{는 답에 대해서만 } 1 \text{이고 나머지는 모두 } 0 \\ &= u(\hat{y} - y) \end{aligned}$$

(c) $J(v_c, o, U)$ 은 outside word vector, Uw 's에 대해 미분해라
($w=0$ 일 때와 $w \neq 0$ 일 때가 있음, y, \hat{y}, v_c 로 나타내라)

i) $J(v_c, o, U) = -\log P(O=o | C=c)$ 이다,

$$P(O=o | C=c) = \frac{\exp(U_o^T v_c)}{\sum_w \exp(U_w^T v_c)}$$

$$\text{ii) } J = -\log P(O=o | C=c) = -\log \frac{\exp(U_o^T v_c)}{\sum_w \exp(U_w^T v_c)} = -U_o^T v_c + \log \sum_w \exp(U_w^T v_c)$$

* 미분

iii) $w=0$, outside word가 실제 정답일 경우

$$\begin{aligned} \frac{\partial J}{\partial U_w} &= -v_c + \frac{\partial}{\partial U_w} (\log \sum_w \exp(U_w^T v_c)) = -v_c + \frac{(U_w^T v_c)}{\sum_w \exp(U_w^T v_c)} \cdot \frac{\partial (\sum_w \exp(U_w^T v_c))}{\partial U_w} \\ &= -v_c + P(O=o | C=c) \cdot v_c \\ &= (\hat{y} - y) v_c \quad (P(O=o | C=c) - 1) v_c \end{aligned}$$

iv) $w \neq 0$ outside word가 실제 정답 아닌 경우

$$\begin{aligned} \frac{\partial J}{\partial U_w} &= -v_c + \frac{\partial}{\partial U_w} (U_o^T v_c) + \frac{\partial}{\partial U_w} (\log \sum_w \exp(U_w^T v_c)) \\ &= 0 + \frac{(U_w^T v_c)}{\sum_w \exp(U_w^T v_c)} \cdot \frac{\partial}{\partial U_w} (U_w^T v_c) \\ &= P(O=w | C=c) \cdot v_c \\ &= \hat{y} v_c \end{aligned}$$

(d) 시그모이드 함수

$$\text{Equation (4)} : \sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x + 1}$$

x 가 scalar 일 때, $\sigma(x)$ 를 x 에 대해 미분해라

$$\begin{aligned} \frac{d}{dx} \sigma(x) &= \frac{d}{dx} \left(\frac{1}{1+e^{-x}} \right) = \frac{d}{dx} (1+e^{-x})^{-1} \\ &= -(1+e^{-x})^{-2} (-e^{-x}) \quad (\text{연쇄법칙, 속미분}) \\ &= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{-x}} \cdot \left(\frac{1+e^{-x}-1}{1+e^{-x}} \right) \\ &= \frac{1}{1+e^{-x}} \left(1 - \frac{1}{1+e^{-x}} \right) \\ &= \sigma(x)(1-\sigma(x)) \end{aligned}$$

(e) For negative sample $\{w_1, w_2, \dots, w_k\} \neq 0$,

$$J(v_c, o, u) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^K \log(\sigma(-u_k^T v_c))$$

where $\sigma(\cdot)$ is a sigmoid function. $\sigma(z) = \frac{1}{1+e^{-z}}$

$$\begin{aligned} i) \frac{\partial J}{\partial v_c} &= -\frac{1}{\sigma(u_o^T v_c)} \cdot \frac{\partial}{\partial v_c} \sigma(u_o^T v_c) - \sum_{k=1}^K \frac{1}{\sigma(u_k^T v_c)} \cdot \frac{\partial}{\partial v_c} \sigma(-u_k^T v_c) \\ &= -\frac{1}{\sigma(u_o^T v_c)} \sigma(u_o^T v_c) (1 - \sigma(u_o^T v_c)) \cdot \frac{\partial}{\partial v_c} (u_o^T v_c) \quad (\because \text{시그모이드 미분}) \\ &\quad - \sum_{k=1}^K \frac{1}{\sigma(-u_k^T v_c)} \sigma(-u_k^T v_c) (1 - \sigma(-u_k^T v_c)) \cdot \frac{\partial}{\partial v_c} (-u_k^T v_c) \end{aligned}$$

$$\begin{aligned} &= -(1 - \sigma(u_o^T v_c)) \cdot u_o - \sum_{k=1}^K (1 - \sigma(-u_k^T v_c)) (-u_k) \\ &= (\sigma(u_o^T v_c) - 1) u_o + \sum_{k=1}^K (1 - \sigma(-u_k^T v_c)) u_k \end{aligned}$$

$$\begin{aligned} ii) \frac{\partial J}{\partial u_o} &= -\frac{1}{\sigma(u_o^T v_c)} \cdot \frac{\partial}{\partial u_o} \sigma(u_o^T v_c) - \sum_{k=1}^K \frac{\partial}{\partial u_o} \log \sigma(-u_k^T v_c) \\ &= -\frac{1}{\sigma(u_o^T v_c)} \sigma(u_o^T v_c) (1 - \sigma(u_o^T v_c)) \cdot \frac{\partial}{\partial u_o} (u_o^T v_c) - 0 \\ &= (\sigma(u_o^T v_c) - 1) \cdot v_c \end{aligned}$$

negative sample이므로
 $\{w_1, \dots, w_k\} \neq 0$ 이다.
 그러므로 모든 k 에 대해서
 $\frac{\partial}{\partial u_o} \log \sigma(-u_k^T v_c) = 0$ 이 된다.

* negative sampling의 목적은
 target 단어(0)와 연관성이 높을
 것이라고 추정되는 단어를 뽑는 것
 이다.

$$\begin{aligned} iii) \frac{\partial J}{\partial u_k} &= \frac{\partial}{\partial u_k} (-\log \sigma(u_o^T v_c)) - \sum_{k=1}^K \frac{\partial}{\partial u_k} (\log \sigma(-u_k^T v_c)) \\ &= -\sum_{k=1}^K \frac{1}{\sigma(-u_k^T v_c)} \frac{\partial}{\partial u_k} (\sigma(-u_k^T v_c)) \\ &= -\sum_{k=1}^K \frac{1}{\sigma(-u_k^T v_c)} \sigma(-u_k^T v_c) (1 - \sigma(-u_k^T v_c)) \cdot \frac{\partial}{\partial u_k} (-u_k^T v_c) \\ &= \sum_{k=1}^K (1 - \sigma(-u_k^T v_c)) v_c \end{aligned}$$

negative sample $\{w_1, \dots, w_k\} \neq 0$ 이므로

$-\log \sigma(u_o^T v_c)$ 는 u_k 에 대해 미분하면

상수취급된다. $\frac{\partial}{\partial u_k} (-\log \sigma(u_o^T v_c)) = 0$ 이다.

기존의 방식(naive softmax) 보다 새로운 방식(negative sampling)의 성능이 더 좋은 이유 :

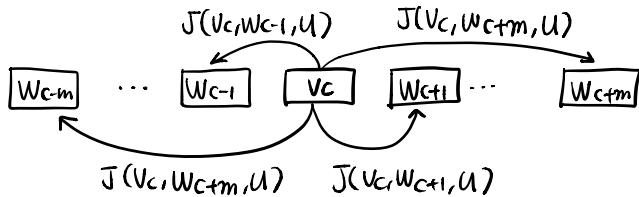
naive softmax 같은 경우는 언제 vocabulary에 대해서 softmax 연산을 해야 한다는 문제가 있다.

전체 단어에 대해서 $u_w^T v_c$ 의 연산 후 지수승을 하고 모두 더해줘야 하는데, 이러한 연산은 비용이 너무 많이 든다는 단점이 있다.

하지만, Negative sampling의 경우, negative sample 몇 개만 뽑아서 해당 negative sample에 대한 loss function을 연산하기 때문에, 상대적으로 연산 비용이 적다는 장점이 있다.

전반적인 해설 필요

$$(f) J_{\text{skip-gram}}(v_c, w_{t-m}, \dots, w_{t+m}, u) = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, u)$$



$$(i) \frac{\partial}{\partial u} J_{\text{S-g}}(v_c, w_{t-m}, \dots, w_{t+m}, u) = \frac{\partial}{\partial u} \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, u)$$

Skip-gram 목적식을
outside vector, u 에 대해서 미분

$$(ii) \frac{\partial}{\partial v_c} J_{\text{S-g}}(v_c, w_{t-m}, \dots, w_{t+m}, u) = \frac{\partial}{\partial v_c} \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, u)$$

center vector, $(v_c)_0$ 에 대해서 미분

$$*(iii) \frac{\partial}{\partial v_w} J_{\text{S-g}}(v_c, w_{t-m}, \dots, w_{t+m}, u) = \frac{\partial}{\partial v_w} \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, u), w \neq c$$