NLP Assignment 2

Parisa Yalsavar

April 4, 2021

1 a

y is a one-hot vector, all elements are zero except for the oth index:

$$y = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -[y_1 \log(\hat{y}_1) + \cdots + y_o \log(\hat{y}_o) + \cdots + y_{|V|} \log(\hat{y}_{|V|})] = -\log(\hat{y}_o)$$

2 b

$$\frac{\partial J}{\partial v_c} = -\frac{\partial u_o^T v_c}{\partial v_c} + \frac{\partial[\log \sum_{w \in Vocab} exp(u_w^T v_c)]}{\partial v_c} = -u_o + \frac{1}{\sum_w exp(u_w^T v_c)} \sum_x u_x^T exp(u_x^T v_c) =$$

$$-u_o + \sum_x \frac{exp(u_x^T v_c)}{\sum_w exp(u_w^T v_c)} u_x = -u_o + \sum_x p(x|c) u_x = -u_o + \sum_x \hat{y}_x u_x = U(y - \hat{y})$$

3 c

$w = o$

$$\frac{\partial J}{\partial u_o} = -\frac{\partial u_o^T v_c}{\partial u_o} + \frac{\partial[\log \sum_{w \in Vocab} exp(u_w^T v_c)]}{\partial u_o} = -v_c + \frac{1}{\sum_w exp(u_w^T v_c)} \frac{\partial[\sum exp(u_x^T v_c)]}{\partial u_o}$$

$$= -v_c + \frac{1}{\sum_w exp(u_w^T v_c)} [v_c e^{u_o^T v_c}] = -v_c + v_c \frac{e^{u_o^T v_c}}{\sum_w exp(u_w^T v_c)} = v_c(\hat{y}_o - 1)$$

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

$w \neq o$

$$\frac{\partial J}{\partial u_o} = -\frac{\partial u_o^T v_c}{\partial u_o} + \frac{\partial[\log \sum_{w \in Vocab} exp(u_w^T v_c)]}{\partial u_o} = 0 + \frac{1}{\sum_w exp(u_w^T v_c)} \frac{\partial[\sum exp(u_x^T v_c)]}{\partial u_o}$$

$$= 0 + \frac{1}{\sum_w exp(u_w^T v_c)} [v_c e^{u_w^T v_c}] = 0 + v_c \frac{e^{u_w^T v_c}}{\sum_w exp(u_w^T v_c)} = v_c(\hat{y}_{w \neq o})$$

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

4 d

$$\frac{\partial J(v_c,o,U)}{\partial U} = \frac{\partial J(v_c,o,U)}{\partial u_1} + \frac{\partial J(v_c,o,U)}{\partial u_2} + \cdots + \frac{\partial J(v_c,o,U)}{\partial u_{|Vocab|}}$$

5 e

$$\begin{aligned}\frac{\partial \sigma(x)}{\partial x} &= \frac{e^x(e^x+1) - e^x(e^x)}{(e^x+1)^2}\\ &= \frac{e^x}{(1+e^x)^2}\\ &= (\frac{1}{e^x+1})(\frac{e^x}{1+e^x})\\ &= \sigma(x)(1-\sigma(x))\end{aligned}$$

6 f

In Naive Softmax Loss, all words of the vocabulary are taken into account ($\mathcal{O}(|V|)$), while in Negative Sampling Loss only K samples are taken into account ($\mathcal{O}(|K|)$).

6.1 v_c

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

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

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

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

6.2 u_o

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

6.3 u_k

$$
\begin{aligned}
\frac{\partial J}{\partial u_k} &= -\frac{\partial[\log(\sigma(u_o^T v_c))]}{\partial u_k} - \frac{\partial \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))}{\partial u_k} \\
&= 0 - \frac{\partial[\log(\sigma(-u_1^T v_c)) + \cdots + \log(\sigma(-u_k^T v_c))]}{\partial u_k} \\
&= -\frac{1}{\sigma(-u_k^T v_c)} \sigma(-u_k^T v_c)(1 - \sigma(-u_k^T v_c))(-v_c) \\
&= v_c(1 - \sigma(-u_k^T v_c))
\end{aligned}
$$

7 g

$$
\begin{aligned}
\frac{\partial J}{\partial u_k} &= -\frac{\partial[\log(\sigma(u_o^T v_c))]}{\partial u_k} - \frac{\partial \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))}{\partial u_k} \\
&= 0 - \frac{\partial[\log(\sigma(-u_1^T v_c)) + \cdots + \log(\sigma(-u_k^T v_c))]}{\partial u_k} \\
&= -\frac{k}{\sigma(-u_k^T v_c)} \sigma(-u_k^T v_c)(1 - \sigma(-u_k^T v_c))(-v_c) \\
&= k v_c(1 - \sigma(-u_k^T v_c))
\end{aligned}
$$

8 h

(i)

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

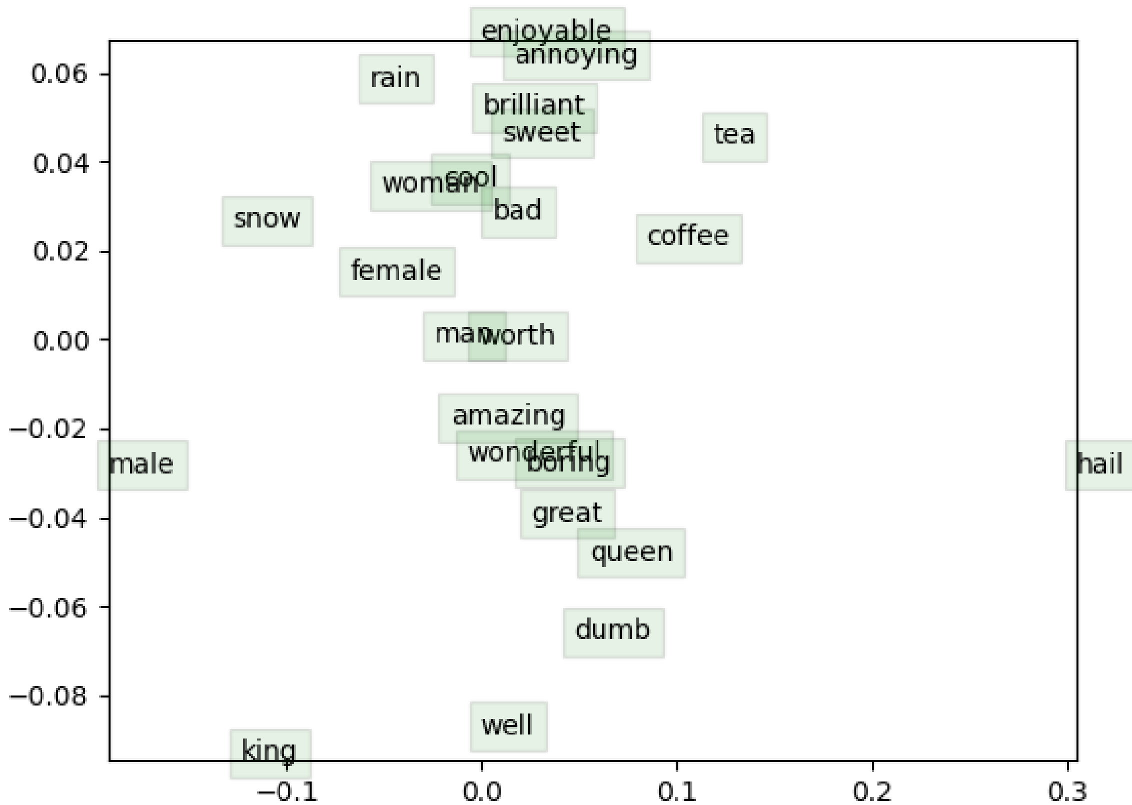
(ii)

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

(iii)

$$
\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \cdots, w_{t+m})}{\partial v_w}_{(w \ne c)} = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \frac{J(v_c, w_{t+j}, U)}{\partial v_w}_{(w \ne c)} = 0
$$

9 Coding



There are two types of clusters: 1.synonym clusters: like *amazing, wonderful, great* . Because these words can be used interchangeably they are clustered together. 2. antonym clusters: words that have opposite meanings but since they have been seen together in the text, they are close to each other like *enjoyable, annoying*. Since word embeddings have high dimenstions and we have compressed some word vectors and lost a lot of information to get this 2D picture, some words are put close together while they don't necessarily have a relation like *woman, cool, bad*.