Assignment 2

20182011 Vyacheslav Kim

April 23, 2023

a) Since y is a one-hot vector with a 1 for the true outside word o, and 0 everywhere else, thus, $w = o$ and $-\sum_{w \in \text{Vocab}} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$.

b)

$$\begin{aligned} \frac{\partial}{\partial v_c} J_{\text{naive-softmax}}(v_c, o, U) &= \frac{\partial}{\partial v_c} - \log(\hat{y}_o) = \frac{\partial}{\partial v_c} - \log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in \text{Vocab}} \exp(u_w^T v_c)}\right) \\ &= \frac{\partial}{\partial v_c} - u_o^T v_c + \log \sum_{w \in \text{Vocab}} \exp(u_w^T v_c) = -u_o^T + \sum_{w \in \text{Vocab}} \frac{\exp(u_w^T v_c) u_w^T}{\sum_{w \in \text{Vocab}} \exp(u_w^T v_c)} \\ &= -u_o^T + \sum_{w \in \text{Vocab}} \hat{y}_w \cdot u_w^T = U\hat{y} - Uy = U(\hat{y} - y) \end{aligned} \tag{1}$$

1) Gradient is zero when $\hat{y} = y$ which means that our predicted distribution is perfectly aligned with ground truth.
2) Subtracting the gradient updates v_c in the direction that minimizes the loss and improves its ability to predict a next outside word.

c) When $w = o$

$$\begin{aligned} \frac{\partial}{\partial u_o} J_{\text{naive-softmax}}(v_c, o, U) &= \frac{\partial}{\partial u_o} - \log\left(\frac{\exp(u_0^T v_c)}{\sum_{w \in \text{Vocab}} \exp(u_w^T v_c)}\right) \\ &= \frac{\partial}{\partial u_o} - u_0^T v_c + \log \sum_{w \in \text{Vocab}} \exp(u_w^T v_c) = -v_c + \frac{\exp(u_0^T v_c) v_c}{\sum_{w \in \text{Vocab}} \exp(u_w^T v_c)} \\ &\hat{y}_o v_c - v_c = (\hat{y}_o - 1) v_c \end{aligned} \tag{2}$$

When $w \neq o$

$$\begin{aligned} \frac{\partial}{\partial u_w} J_{\text{naive-softmax}}(v_c, o, U) &= \frac{\partial}{\partial u_w} - \log\left(\frac{\exp(u_0^T v_c)}{\sum_{w \in \text{Vocab}} \exp(u_w^T v_c)}\right) \\ &= \frac{\partial}{\partial u_w} - u_0^T v_c + \log \sum_{w \in \text{Vocab}} \exp(u_w^T v_c) = \frac{\exp(u_w^T v_c) v_c}{\sum_{w \in \text{Vocab}} \exp(u_w^T v_c)} = \hat{y}_w v_c \end{aligned} \tag{3}$$

d)

$$\begin{aligned} \frac{\partial}{\partial U} J_{\text{naive-softmax}}(v_c, o, U) &= [\frac{\partial}{\partial u_1} J(v_c, o, U), \frac{\partial}{\partial u_2} J(v_c, o, U), \ldots, \frac{\partial}{\partial u_{|\text{Vocab}|}} J(v_c, o, U)] \\ &= [(\hat{y}_1 - y_1) v_c, (\hat{y}_2 - y_2) v_c, \ldots, (\hat{y}_{|\text{Vocab}|} - y_{|\text{Vocab}|}) v_c,] \end{aligned} \tag{4}$$

e)

$$f'(x) = \begin{cases} f'(x) = 1 & \text{if } x > 0 \\ f'(x) = 0 & \text{if } x < 0 \end{cases} \tag{5}$$

f)

$$\begin{aligned}\sigma'(x) = \frac{d}{dx}\frac{e^x}{e^x+1} = \frac{(e^x+1)e^x - e^2x}{(e^x+1)^2} = \frac{e^x}{(e^x+1)^2} = \sigma(x)\cdot\frac{e^x-e^x+1}{e^x+1} \\ = \sigma(x)\left(\frac{e^x+1}{e^x+1} - \frac{e^x}{e^x+1}\right) = \sigma(x)(1-\sigma(x))\end{aligned} \tag{6}$$

g) i)

$$\begin{aligned}\frac{\partial}{\partial v_c}J_{\text{neg-sample}}(v_c,o,U) = \frac{\partial}{\partial v_c} - \log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K}\log(\sigma(-u_{w_s}^T v_c)) \\ = \frac{\sigma(u_o^T v_c)(\sigma(u_o^T v_c)-1)u_o^T}{\sigma(u_o^T v_c)} + \sum_{s=1}^{K}\frac{\sigma(-u_{w_s}^T v_c)(1-\sigma(-u_{w_s}^T v_c))u_{w_s}^T}{\sigma(-u_{w_s}^T v_c)} \\ = (\sigma(u_o^T v_c)-1)u_o^T + \sum_{s=1}^{K}\sigma(u_{w_s}^T v_c)u_{w_s}^T\end{aligned} \tag{7}$$

Note:

$$1-\sigma(-x) = 1 - \frac{1}{1+e^x} = \frac{e^x}{1+e^x} = \sigma(x) \tag{8}$$

$$\begin{aligned}\frac{\partial}{\partial u_o}J_{\text{neg-sample}}(v_c,o,U) = \frac{\partial}{\partial u_o} - \log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K}\log(\sigma(-u_{w_s}^T v_c)) \\ = \frac{\sigma(u_o^T v_c)(\sigma(u_o^T v_c)-1)v_c}{\sigma(u_o^T v_c)} = (\sigma(u_o^T v_c)-1)v_c\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial}{\partial u_{w_s}}J_{\text{neg-sample}}(v_c,o,U) = \frac{\partial}{\partial u_{w_s}} - \log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K}\log(\sigma(-u_{w_s}^T v_c)) \\ = \frac{\sigma(-u_{w_s}^T v_c)(1-\sigma(-u_{w_s}^T v_c))v_c}{\sigma(-u_{w_s}^T v_c)} = \sigma(u_{w_s}^T v_c)v_c\end{aligned} \tag{10}$$

ii) $\sigma(u_o^T v_c) - 1$ and $\sigma(-u_{w_s}^T v_c) - 1$ can be reused to compute partial derivatives with respect to v_c, u_o and u_{w_s}.

$$\sigma(U_{o,\{w1,\ldots,wK\}}v_c^T) - \mathbf{1} = [\sigma(u_o^T v_c) - 1, \sigma(-u_{w_1}^T v_c) - 1, \ldots, \sigma(-u_{w_K}^T v_c) - 1] \tag{11}$$

iii) The negative sampling loss function is more efficient than the naive-softmax loss because it only requires computing the gradients for a small number of negative samples instead of the entire vocabulary.

h)

$$\frac{\partial}{\partial u_{w_s}}J_{\text{neg-sample}}(v_c,o,U) = \sum_{\substack{1\le x\le K\\ w_x = w_s}}\sigma(u_{w_x}^T v_c)v_c \tag{12}$$

i)

$$\frac{\partial}{\partial U}J_{\text{skip-gram}}(v_c,w_{t-m},\ldots,w_{t+m},U) = \sum_{\substack{-m\le j\le +m\\ j\neq 0}}\frac{\partial}{\partial U}J_{\text{skip-gram}}(v_c,w_{t+j},U) \tag{13}$$

$$\frac{\partial}{\partial v_c}J_{\text{skip-gram}}(v_c,w_{t-m},\ldots,w_{t+m},U) = \sum_{\substack{-m\le j\le +m\\ j\neq 0}}\frac{\partial}{\partial v_c}J_{\text{skip-gram}}(v_c,w_{t+j},U) \tag{14}$$

$$\frac{\partial}{\partial u_w}J_{\text{skip-gram}}(v_c,w_{t-m},\ldots,w_{t+m},U) = \sum_{\substack{-m\le j\le +m\\ j\neq 0}}\frac{\partial}{\partial u_w}J_{\text{skip-gram}}(v_c,w_{t+j},U), \text{ when } w\neq c \tag{15}$$

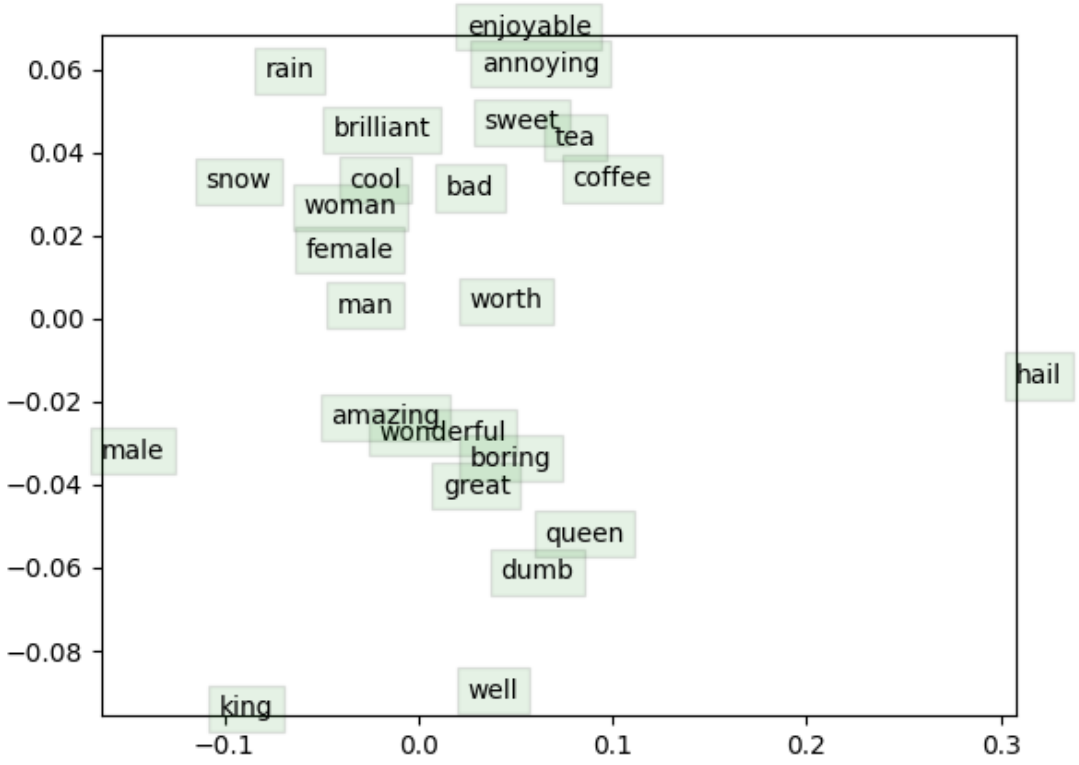

Figure 1: Word vectors

2) In figure 2, we can observe that opposite meaning words are located closely: man and woman, enjoyable and annoying, wonderful and boring, cool and bad. Similar meaning words also located closely: woman and man, tea and coffee, rain and snow, amazing and wonderful.