CS 224n Assignment #2 Written Part

(a)

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

$$= -[0 \cdot log(\hat{y}_1) + \cdots + y_o log(\hat{y}_o) + \cdots + 0 \cdot log(\hat{y}_w)]$$

$$= -1 \cdot log(\hat{y}_o) = -log(\hat{y}_o)$$

(b)

$$\frac{\partial \boldsymbol{J}_{naive-softmax}(\boldsymbol{v_c}, o, \boldsymbol{U})}{\partial \boldsymbol{v_c}} = -\frac{\partial log(\hat{y}_o)}{\partial \boldsymbol{v_c}} = -\frac{\partial}{\partial \boldsymbol{v_c}} log(\frac{exp(\boldsymbol{u_o}^T \boldsymbol{v_c})}{\sum_{w \in Vocab} exp(\boldsymbol{u_w}^T \boldsymbol{v_c})})$$

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

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

$$= (\sum_{w \in Vocab} \hat{y_w} \boldsymbol{u_w}) - \boldsymbol{u_o}$$

$$= (\hat{y_1} \boldsymbol{u_1} - 0 \cdot \boldsymbol{u_1}) + \cdots + (\hat{y_o} \boldsymbol{u_o} - 1 \cdot \boldsymbol{u_o}) + \cdots + (\hat{y_w} \boldsymbol{u_w} - 0 \cdot \boldsymbol{u_w})$$

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

(c)

$$\frac{\partial \boldsymbol{J}_{naive-softmax}(\boldsymbol{v_c}, o, \boldsymbol{U})}{\partial \boldsymbol{u_w}} = -\frac{\partial log(\hat{y}_o)}{\partial \boldsymbol{u_w}} = -\frac{\partial}{\partial \boldsymbol{u_w}} log(\frac{exp(\boldsymbol{u_o}^T \boldsymbol{v_c})}{\sum_{w \in Vocab} exp(\boldsymbol{u_w}^T \boldsymbol{u_c})})$$

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

When $w = o$**,**

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

$$= \hat{y_o} \boldsymbol{v_c} - \boldsymbol{v_c}$$

When $w \neq o$**,**

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

$$= (\sum_{w \in Vocab, w \neq o} \hat{y_w} \boldsymbol{v_c}) - 0 = \sum_{w \in Vocab, w \neq o} \hat{y_w} \boldsymbol{v_c}$$

Therefore,

$$\frac{\partial \boldsymbol{J}_{naive-softmax}(\boldsymbol{v_c}, o, \boldsymbol{U})}{\partial \boldsymbol{u_w}} = (\sum_{w \in Vocab} \hat{y_w} \boldsymbol{v_c}) - \boldsymbol{v_c}$$

$$= (\hat{y_1} \boldsymbol{v_c} - 0 \cdot \boldsymbol{v_c}) + \cdots + (\hat{y_o} \boldsymbol{v_c} - 1 \cdot \boldsymbol{v_c}) + \cdots + (\hat{y_w} \boldsymbol{v_c} - 0 \cdot \boldsymbol{v_c})$$

$$= \boldsymbol{v_c} (\hat{\boldsymbol{y}} - \boldsymbol{y})^T$$

(d)

$$\frac{d\sigma(\boldsymbol{x})}{d\boldsymbol{x}} = \frac{d}{d\boldsymbol{x}} \frac{e^{\boldsymbol{x}}}{e^{\boldsymbol{x}} + 1} = 0 - \frac{d}{d\boldsymbol{x}} \frac{1}{e^{\boldsymbol{x}} + 1} = \frac{e^{\boldsymbol{x}}}{(e^{\boldsymbol{x}} + 1)^2}$$

(e)

$$\frac{\frac{d\sigma(\boldsymbol{x})}{d\boldsymbol{x}}}{\sigma(\boldsymbol{x})} = \frac{\frac{e^{\boldsymbol{x}}}{(e^{\boldsymbol{x}}+1)^2}}{\frac{e^{\boldsymbol{x}}}{e^{\boldsymbol{x}}+1}} = \frac{1}{e^{\boldsymbol{x}} + 1} = \sigma(-\boldsymbol{x})$$

$$\frac{\partial \boldsymbol{J}_{negative-sample}(\boldsymbol{v_c}, o, \boldsymbol{U})}{\partial \boldsymbol{v_c}} = -\frac{\partial}{\partial \boldsymbol{v_c}} log(\sigma(\boldsymbol{u_o}^T \boldsymbol{v_c})) - \frac{\partial}{\partial \boldsymbol{v_c}} \sum_{k=1}^{K} log(\sigma(-\boldsymbol{u_k}^T \boldsymbol{v_c}))$$

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

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

$$\frac{\partial \boldsymbol{J}_{negative-sample}(\boldsymbol{v_c}, o, \boldsymbol{U})}{\partial \boldsymbol{u_o}} = -\frac{\partial}{\partial \boldsymbol{u_o}} log(\sigma(\boldsymbol{u_o}^T \boldsymbol{v_c})) - \frac{\partial}{\partial \boldsymbol{u_o}} \sum_{k=1}^{K} log(\sigma(-\boldsymbol{u_k}^T \boldsymbol{v_c}))$$

$$= -\frac{\frac{\partial}{\partial \boldsymbol{u_o}} \sigma(\boldsymbol{u_o}^T \boldsymbol{v_c})}{\sigma(\boldsymbol{u_o}^T \boldsymbol{v_c})} - 0 = -\sigma(-\boldsymbol{u_o}^T \boldsymbol{v_c}) \cdot \boldsymbol{v_c}$$

$$\frac{\partial \boldsymbol{J}_{negative-sample}(\boldsymbol{v_c}, o, \boldsymbol{U})}{\partial \boldsymbol{u_k}} = -\frac{\partial}{\partial \boldsymbol{u_k}} log(\sigma(\boldsymbol{u_o}^T \boldsymbol{v_c})) - \frac{\partial}{\partial \boldsymbol{u_k}} \sum_{k=1}^{K} log(\sigma(-\boldsymbol{u_k}^T \boldsymbol{v_c}))$$

$$= -0 - 0 - \cdots - \frac{\frac{\partial}{\partial \boldsymbol{u_k}} \sigma(-\boldsymbol{u_k}^T \boldsymbol{v_c})}{\sigma(-\boldsymbol{u_k}^T \boldsymbol{v_c})} - 0 - \cdots - 0 = \sigma(\boldsymbol{u_k}^T \boldsymbol{v_c}) \cdot \boldsymbol{v_c}$$

It is much more efficient because with the aid of some tricks working out the derivative of sigmoid function is much easier than handling naive-softmax function.

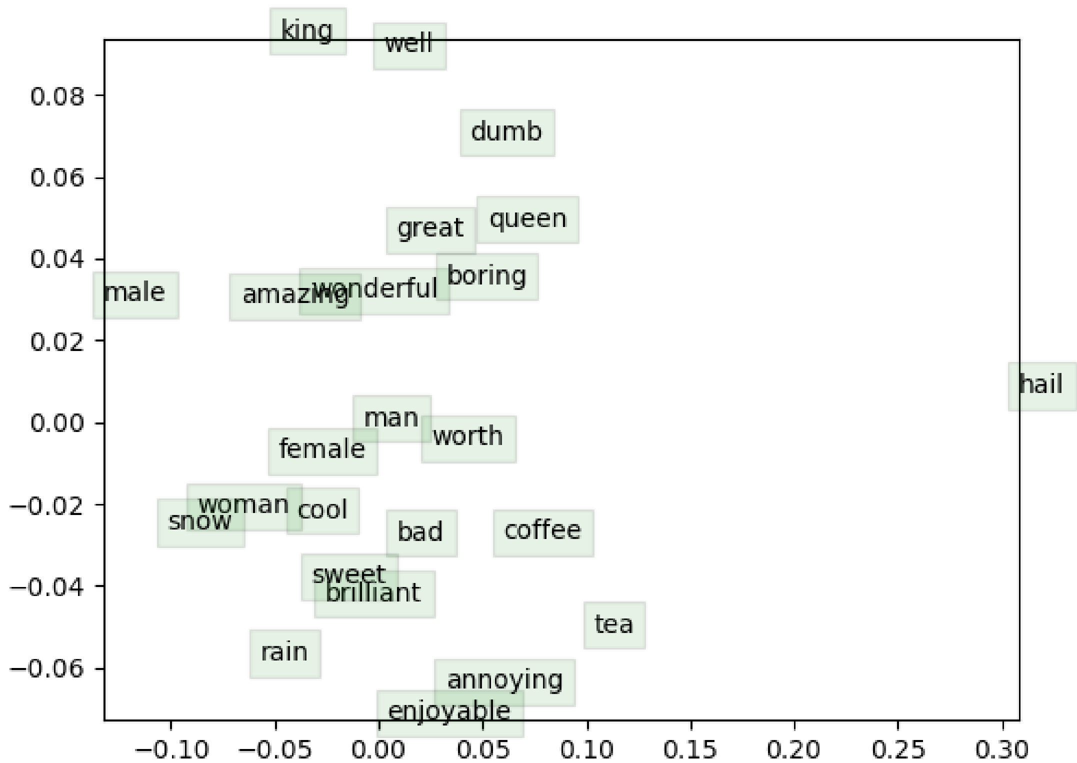
(f)

$$\frac{\partial \boldsymbol{J}_{skip-gram}(\boldsymbol{v_c}, w_{t-m}, \cdots, w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{U}} = \sum_{-m<j<m, j\neq 0} \frac{\partial \boldsymbol{J}_{skip-gram}(\boldsymbol{v_c}, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{U}}$$

$$\frac{\partial \boldsymbol{J}_{skip-gram}(\boldsymbol{v_c}, w_{t-m}, \cdots, w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v_c}} = \sum_{-m<j<m, j\neq 0} \frac{\partial \boldsymbol{J}_{skip-gram}(\boldsymbol{v_c}, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{v_c}}$$

$$\frac{\partial \boldsymbol{J}_{skip-gram}(\boldsymbol{v_c}, w_{t-m}, \cdots, w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v_w}} = 0$$

[My Word Vector]



Some words describing feelings or emotions like "enjoyable" and "annoying", "amazing" and "wonderful" cluster together. If we have the word vector of "male" minus "female" and the result will approximately equal the word vector of "king" minus "queen".