CS224 Winter 2019 Assignment 2

Name: Dat Nguyen

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

1 Written: Understanding word2vec (23 points)

(a) We have

$$
\begin{aligned}
-\sum_{w\in\text{Vocab}} y_w \log(\hat{y}_w) &= -\sum_{w\in\text{Vocab}} \mathbb{1}\{w=o\}\log(\hat{y}_w) \\
&= -\log\hat{y}_o
\end{aligned}
$$

(b) We have

$$
\begin{aligned}
\mathbf{J}_{\text{naive-softmax}} &= -\log\frac{\exp(\mathbf{u}_o^T\mathbf{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\mathbf{u}_w^T\mathbf{v}_c)} \\
&= -\mathbf{u}_o^T\mathbf{v}_c + \log\sum_{w\in\text{Vocab}}\exp(\mathbf{u}_w^T\mathbf{v}_c)
\end{aligned}
$$

So

$$
\begin{aligned}
\frac{\partial \mathbf{J}_{\text{naive-softmax}}}{\partial \mathbf{v}_c} &= -\mathbf{u}_o + \frac{1}{\sum_{w\in\text{Vocab}}\exp(\mathbf{u}_w^T\mathbf{v}_c)}\frac{\partial\sum_{w\in\text{Vocab}}\exp(\mathbf{u}_w^T\mathbf{v}_c)}{\partial\mathbf{v}_c} \\
&= -\mathbf{u}_o + \sum_{w\in\text{Vocab}}\mathbf{u}_w\frac{\exp(\mathbf{u}_w^T\mathbf{v}_c)}{\sum_{w'\in\text{Vocab}}\exp(\mathbf{u}_{w'}^T\mathbf{v}_c)} \\
&= -\mathbf{u}_o + \sum_{w\in\text{Vocab}}\mathbf{u}_w\hat{\mathbf{y}}_w
\end{aligned}
$$

(c) We have

$$
\begin{aligned}
\frac{\partial \mathbf{J}_{\text{naive-softmax}}}{\partial \mathbf{u}_w} &= -\mathbb{1}\{w=o\}\mathbf{v}_c + \mathbf{v}_c\frac{\exp(\mathbf{u}_w^T\mathbf{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\mathbf{u}_w^T\mathbf{v}_c)} \\
&= -\mathbf{y}_w\mathbf{v}_c + \mathbf{v}_c\hat{\mathbf{y}}_w \\
&= \mathbf{v}_\mathbf{c}(\hat{\mathbf{y}}_w - \mathbf{y}_w)
\end{aligned}
$$

(d) We have

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

(e) Derivative with respect to $\mathbf{v}_c$

$$\begin{aligned}\frac{\partial\mathbf{J}_{\text{neg-sample}}(\mathbf{v}_c,o,\mathbf{U})}{\partial\mathbf{v}_c} &= -\frac{\sigma(\mathbf{u}_o^T\mathbf{v}_c)(1-\sigma(\mathbf{u}_o^T\mathbf{v}_c))\mathbf{u}_o}{\sigma(\mathbf{u}_o^T\mathbf{v}_c)} - \sum_{k=1}^{K}\frac{\sigma(-\mathbf{u}_k^T\mathbf{v}_c)(1-\sigma(-\mathbf{u}_k^T\mathbf{v}_c))(-\mathbf{u}_k)}{\sigma(-\mathbf{u}_k^T\mathbf{v}_c)}\\ &= (\sigma(\mathbf{u}_o^T\mathbf{v}_c)-1)\mathbf{u}_o + \sum_{k=1}^{K}(1-\sigma(-\mathbf{u}_k^T\mathbf{v}_c))\mathbf{u}_k\end{aligned}$$

Derivative with respect to $\mathbf{u}_o$

$$\begin{aligned}\frac{\partial\mathbf{J}_{\text{neg-sample}}(\mathbf{v}_c,o,\mathbf{U})}{\partial\mathbf{u}_o} &= -\frac{\sigma(\mathbf{u}_o^T\mathbf{v}_c)(1-\sigma(\mathbf{u}_o^T\mathbf{v}_c))\mathbf{v}_c}{\sigma(\mathbf{u}_o^T\mathbf{v}_c)}\\ &= (\sigma(\mathbf{u}_o^T\mathbf{v}_c)-1)\mathbf{v}_c\end{aligned}$$

Derivative with respect to $\mathbf{u}_k$

$$\begin{aligned}\frac{\partial\mathbf{J}_{\text{neg-sample}}(\mathbf{v}_c,o,\mathbf{U})}{\partial\mathbf{u}_k} &= -\frac{\sigma(-\mathbf{u}_k^T\mathbf{v}_c)(1-\sigma(-\mathbf{u}_k^T\mathbf{v}_c))(-\mathbf{v}_c)}{\sigma(-\mathbf{u}_k^T\mathbf{v}_c)}\\ &= (1-\sigma(-\mathbf{u}_k^T\mathbf{v}_c))\mathbf{v}_c\end{aligned}$$

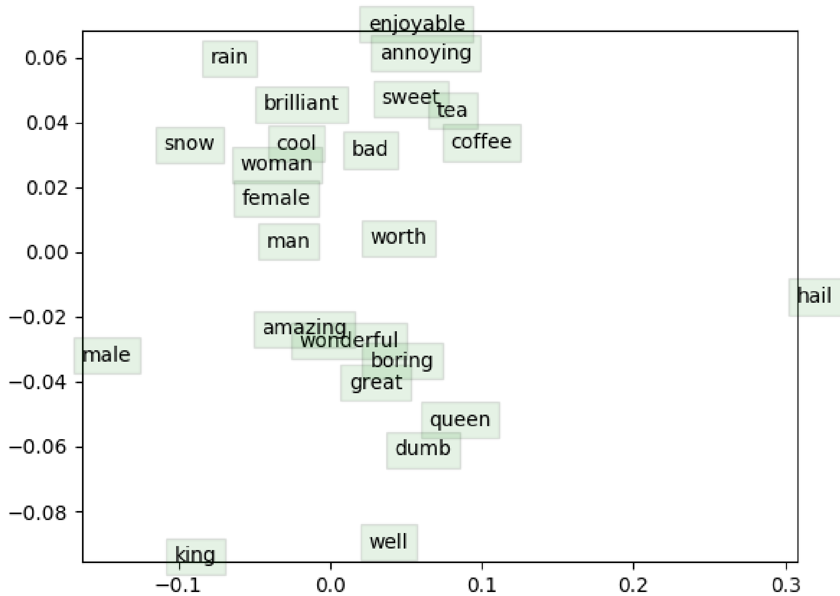
(f) We have

$$\frac{\partial\mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c,w_{t-m},\dots,w_{t+m},\mathbf{U})}{\partial\mathbf{U}} = \sum_{\substack{-m\le j\le m\\ j\ne 0}}\frac{\partial\mathbf{J}(\mathbf{v}_c,\mathbf{w}_{t+j},\mathbf{U})}{\partial\mathbf{U}}$$

$$\frac{\partial\mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c,w_{t-m},\dots,w_{t+m},\mathbf{U})}{\partial\mathbf{v}_c} = \sum_{\substack{-m\le j\le m\\ j\ne 0}}\frac{\partial\mathbf{J}(\mathbf{v}_c,\mathbf{w}_{t+j},\mathbf{U})}{\partial\mathbf{v}_c}$$

$$\frac{\partial\mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c,w_{t-m},\dots,w_{t+m},\mathbf{U})}{\partial\mathbf{v}_w} = \sum_{\substack{-m\le j\le m\\ j\ne 0}}\frac{\partial\mathbf{J}(\mathbf{v}_c,\mathbf{w}_{t+j},\mathbf{U})}{\partial\mathbf{v}_w}$$

2 Coding: Implementing word2vec (20 points)



I can see that some words appear in similar context are close together such as "enjoyable" and "annoying", "tea" and "coffee", "woman" and "female", "amazing" and "wonderful" and "boring" and "great. Also, if we subtract the vector of words "king" to "male", the resulting vector is similar as we subtract vectors for "queen" and "female".