NLP
Assignment 2

Alireza Moradi

April 3, 2021

1 Written

a

$\boldsymbol{y}$ is a one-hot encoded vector, so all of the elements are $\mathbf{0}$ except $\boldsymbol{y_o}$ which is $\mathbf{1}$ so the sum in $\boldsymbol{-\sum_{w\in Vocab} y_w \log(\hat{y}_w)}$ will be simplified as $\boldsymbol{-\log(\hat{y}_o)}$.

b

$$\left.\begin{aligned} \hat{\boldsymbol{y}} = P(O=o \mid C=c) = \tfrac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \\ \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log P(O=o \mid C=c) \end{aligned}\right\} \Rightarrow$$

$$\begin{aligned} \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= -\log\left(\frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}\right) \\ &= -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log \sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \end{aligned}$$

And So:

$$\begin{aligned} \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} &= \frac{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\boldsymbol{u}_w}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} - \boldsymbol{u}_o \\ &= \sum_{w\in\text{Vocab}}\left(\frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}\boldsymbol{u}_w\right) - \boldsymbol{u}_o \\ &= \sum_{w\in\text{Vocab}} (\hat{\boldsymbol{y}}\boldsymbol{u}_w) - \boldsymbol{u}_o \\ &= \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}) \end{aligned}$$

c

$$\begin{aligned} \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_w} &= -\boldsymbol{v}_c\boldsymbol{y}_w + \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_{\boldsymbol{w}}^\top \boldsymbol{v}_c)}\frac{\partial(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\partial \boldsymbol{u}_w} \\ &= -\boldsymbol{v}_c\boldsymbol{y}_w + \hat{\boldsymbol{y}}_w\boldsymbol{v}_c \\ &= \begin{cases} \boldsymbol{v}_c(\hat{\boldsymbol{y}}_w - \boldsymbol{y}_w) & w = o \\ 0 & w \neq o \end{cases} \end{aligned}$$

d

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{U}} = \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_1} + \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_2} + ... + \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_{|Vocab|}}$$

e

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

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

f

$$\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \sum_{k=1}^{K} \log(\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c))$$

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

$$\frac{\partial \boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = -\boldsymbol{v}_c(1-\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c))$$

$$\frac{\partial \boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_k} = \boldsymbol{v}_c(1-\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c))$$

In this method we are computing sum over only K negative samples instead of the whole vocabulary and therefor it is much faster.

g

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_k} &= 0 - \frac{k\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)(1-\sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c))(-\boldsymbol{v}_c)}{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)} \\
&= k\boldsymbol{v}_c(1-\sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c))
\end{aligned}$$

h

$$\frac{\partial \boldsymbol{J}_{\text{skip-gram}}\left(\boldsymbol{v}_c, w_{t-m}, ...w_{t+1}, \boldsymbol{U}\right)}{\partial \boldsymbol{U}} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial \boldsymbol{J}\left(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U}\right)}{\partial \boldsymbol{U}}$$

$$\frac{\partial \boldsymbol{J}_{\text{skip-gram}}\left(\boldsymbol{v}_c, w_{t-m}, ...w_{t+1}, \boldsymbol{U}\right)}{\partial \boldsymbol{v}_c} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial \boldsymbol{J}\left(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U}\right)}{\partial \boldsymbol{v}_c}$$

$$\frac{\partial \boldsymbol{J}_{\text{skip-gram}}\left(\boldsymbol{v}_c, w_{t-m}, ...w_{t+1}, \boldsymbol{U}\right)}{\partial \boldsymbol{v}_w} = 0$$

2 Coding

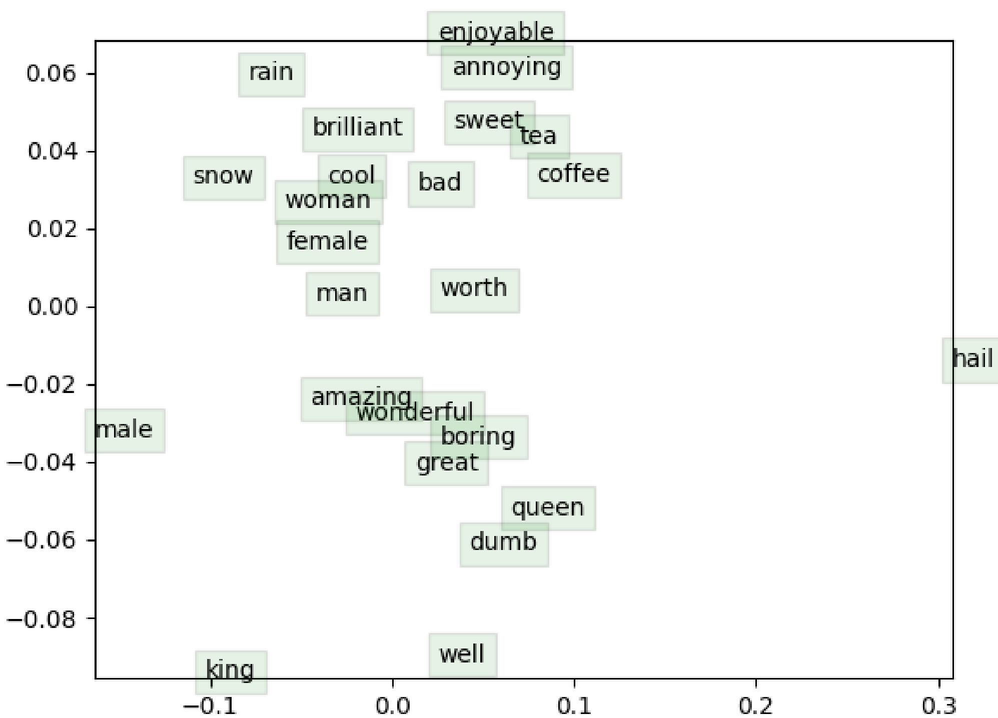

Figure 1: Word vectors training results

We can see that some antonyms or words that have a bold specification that's on the opposite side, close together like `wonderful` and `boring` or `cool` and `bad` or `sweet` and `tea` and `coffee`. Some words are synonyms or have a specification same as the other word, like `brilliant` and `cool` or `amazing` and `wonderful` or `snow` and `rain`. Also some things are not quite acceptable, like `male` which is far from `woman` and `female` and `man`, or `queen` is close to `dumb` and far from `king` which might because we reduced the dimsionality to 2 to be able to visualize it and that might have broken some things.