1 Written: understanding word2vec

(a)

Since $y_w = 1$ if and only if $w = o$, we have

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$$

(b)

$$\boldsymbol{J}_{\text{naive-softmax}} = -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log(\sum_w \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c))$$

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{v}_c} &= -\boldsymbol{u}_o + \sum_{w_0} \frac{\exp(\boldsymbol{u}_{w_0}^\top \boldsymbol{v}_c)}{\sum_w \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \boldsymbol{u}_{w_0} \\ &= -\boldsymbol{U}\boldsymbol{y} + \boldsymbol{U}\hat{\boldsymbol{y}}\end{aligned}$$

(c)

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_k} &= -\boldsymbol{v}_c 1_{k=o} + \frac{\exp(\boldsymbol{u}_k^\top \boldsymbol{v}_c)}{\sum_w \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \boldsymbol{v}_c \\ &= \boldsymbol{v}_c \cdot (\hat{\boldsymbol{y}} - \boldsymbol{y})^\top\end{aligned}$$

(d)

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

(e)

Here we use notation

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

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

(f)

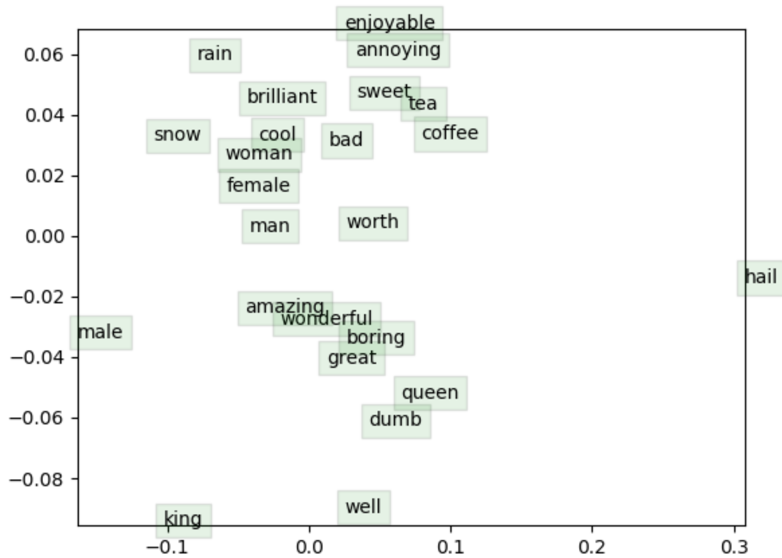
Just sum of gradients of each $\boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$. So I will just skip it

2 Coding: Implementing word2vec

I want to complain that in the previous theoretical part all vectors are row vectors but in the coding part all vectors are row vectors. I know this agrees with the embedding layers but I still want to complain

(c)

My final training loss is about 9.7 and below is my plot



In the plot we can see there are two types of 'clusters' : the words are similar and the words which can be used to replace the other but at the same time change the meaning of the sentence totally