cs224n Assignment #2: word2vec

1. Understanding word2vec

(a) Since the true empirical distribution $\boldsymbol{y}$ has a distribution such that

$$y_w = \begin{cases} 1, & \text{for } w = o \\ 0, & \text{elsewhere} \end{cases},$$

$$\begin{aligned} \text{cross-entropy loss} &= -\sum_{w \in V} y_w \log(\hat{y}_w) \\ &= -y_o \log(\hat{y}_o) - \sum_{\substack{w \in V \\ w \neq o}} y_w \log(\hat{y}_w) \\ &= -y_o \log(\hat{y}_o) = -\log(\hat{y}_o) \\ &= -\log P(O = o | C - c) = \boldsymbol{J}_{\text{naive-softmax}} \end{aligned}$$

(b) Since $\boldsymbol{J}_{\text{naive-softmax}} = -\log(\hat{y}_o) = -\log \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in V} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} = -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log \sum_{w \in V} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)$,

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

(c) From above, $\boldsymbol{J}_{\text{naive-softmax}} = -\log(\hat{y}_o) = -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log \sum_{w \in V} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)$.

$$\begin{aligned} \therefore \frac{\partial}{\partial \boldsymbol{u}_w} \boldsymbol{J} &= \frac{\partial}{\partial \boldsymbol{u}_w}(-\boldsymbol{u}_o^\top \boldsymbol{v}_c) + \frac{\partial}{\partial \boldsymbol{u}_w} \log \sum_{w \in V} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \\ &= \begin{cases} -\boldsymbol{v}_c + \dfrac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\boldsymbol{v}_c}{\sum_{w \in V} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} = -\boldsymbol{v}_c + \hat{y}_w \boldsymbol{v}_c = (\hat{y}_w - y_w)\boldsymbol{v}_c, & \text{for } w = o \text{ since } y_w = 1 \text{ at } w = o \\ \dfrac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\boldsymbol{v}_c}{\sum_{w \in V} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} = \hat{y}_w \boldsymbol{v}_c, & \text{elsewhere} \end{cases} \\ &= \boldsymbol{v}_c(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top \end{aligned}$$

(d) $\boldsymbol{\sigma}(\boldsymbol{x}) = \dfrac{1}{1 + \exp(-\boldsymbol{x})}$

$$\begin{aligned} \therefore \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{\sigma}(\boldsymbol{x}) &= -\frac{-\exp(-\boldsymbol{x})}{(1 + \exp(-\boldsymbol{x}))^2} = \frac{1}{1 + \exp(-\boldsymbol{x})} \cdot \frac{\exp(-\boldsymbol{x})}{1 + \exp(-\boldsymbol{x})} \\ &= \frac{1}{1 + \exp(-\boldsymbol{x})} \cdot (1 - \frac{1}{1 + \exp(-\boldsymbol{x})}) = \boldsymbol{\sigma}(\boldsymbol{x})(1 - \boldsymbol{\sigma}(\boldsymbol{x})) \end{aligned}$$

and, we can easily get $\dfrac{\boldsymbol{\sigma}'(\boldsymbol{x})}{\boldsymbol{\sigma}(\boldsymbol{x})} = 1 - \boldsymbol{\sigma}(\boldsymbol{x})$.

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

$$\therefore \frac{\partial}{\partial \boldsymbol{v}_c}\boldsymbol{J} = -\frac{\partial}{\partial \boldsymbol{v}_c}\log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \frac{\partial}{\partial \boldsymbol{v}_c}\sum_{k=1}^{K}\log(\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)) = -\frac{\sigma'(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)} - \sum_{k=1}^{K}\frac{\sigma'(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}$$

$$= -\boldsymbol{u}_o(1-\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \sum_{k=1}^{K} -\boldsymbol{u}_k(1-\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c))$$

$$= \boldsymbol{u}_o(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)-1) + \sum_{k=1}^{K}\boldsymbol{u}_k\sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c)$$

$$\frac{\partial}{\partial \boldsymbol{u}_o}\boldsymbol{J} = -\frac{\partial}{\partial \boldsymbol{u}_o}\log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \frac{\partial}{\partial \boldsymbol{u}_o}\sum_{k=1}^{K}\log(\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)) = -\frac{\sigma'(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)} - \sum_{k=1}^{K}\frac{\sigma'(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}$$

$$= -\boldsymbol{v}_c(1-\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) = \boldsymbol{v}_c(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)-1)$$

$$\frac{\partial}{\partial \boldsymbol{u}_k}\boldsymbol{J} = -\frac{\partial}{\partial \boldsymbol{u}_k}\log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \frac{\partial}{\partial \boldsymbol{u}_k}\sum_{k=1}^{K}\log(\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)) = -\frac{\sigma'(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)} - \sum_{k=1}^{K}\frac{\sigma'(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}$$

$$= -\sum_{k=1}^{K} -\boldsymbol{v}_c(1-\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)) = \sum_{k=1}^{K}\boldsymbol{v}_c(1-\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)) = \sum_{k=1}^{K}\boldsymbol{v}_c\sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c)$$

Negative sampling is more efficient because it doesn't need to use all words in vocabulary set to compute the loss but only the fraction of vocabulary set which are used in sampling.

(f) $\boldsymbol{J}_{\text{skip-gram}} = \sum_{\substack{-m\le j\le m\\ j\ne 0}} \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$ when $\boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$ is either $\boldsymbol{J}_{\text{naive-softmax}}$ or $\boldsymbol{J}_{\text{neg-sample}}$

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

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

$$\frac{\partial}{\partial \boldsymbol{v}_w}\boldsymbol{J}_{\text{skip-gram}} = 0 \text{ for } w \ne c$$

2. Implementing word2vec

(c) Some similar words are very close i.e. (amazing, wonderful, great), (woman, female) as expected. But some opposite words are also very close in the figure i.e. (female, man), (enjoyable, annoying). Also, we can find some analogies i.e (female : male :: queen : king) in the figure.

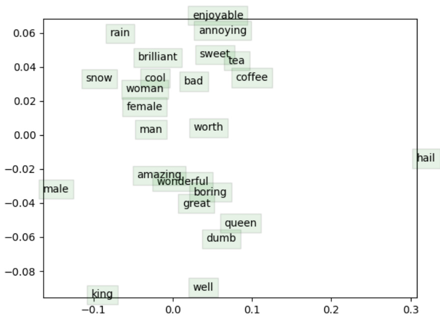

Figure 1: Show time! word vectors