CS224N Natural Language Processing with Deep Learning Assignment #2 Solution

Roy C.K. Chan

In this assignment, let $N = |Vocab|$ be the size of the vocabulary, and D be the dimension of word vectors. There is a subtle difference in defining matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ between the written and coding parts. In the written part, the word vectors are column vectors such that $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{D \times N}$; whereas in the coding part, the word vectors are row vectors so that $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{N \times D}$.

Problem 1(a).

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

Problem 1(b).

$$
\begin{aligned}
\boldsymbol{J} &= \boldsymbol{J}_{naive-softmax}(\boldsymbol{v}_c, o, \boldsymbol{U}) \\
&= -\log(\hat{y}_o) \\
&= -\log \frac{\exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \\
&= -\boldsymbol{u}_o^T \boldsymbol{v}_c + \log \sum_{w \in Vocab} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c) \\
\Longrightarrow \nabla_{\boldsymbol{v}_c} \boldsymbol{J} &= -\boldsymbol{u}_o + \frac{\sum_{w' \in Vocab} \exp(\boldsymbol{u}_{w'}^T \boldsymbol{v}_c) \cdot \boldsymbol{u}_{w'}}{\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \\
&= -\boldsymbol{u}_o + \sum_{w' \in Vocab} \frac{\exp(\boldsymbol{u}_{w'}^T \boldsymbol{v}_c)}{\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \cdot \boldsymbol{u}_{w'} \\
&= -\boldsymbol{u}_o + \sum_{w' \in Vocab} \hat{y}_{w'} \cdot \boldsymbol{u}_{w'} \\
&= -\boldsymbol{U}\boldsymbol{y} + \boldsymbol{U}\hat{\boldsymbol{y}} \\
&= \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}) .
\end{aligned}
$$

Problem 1(c).

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

When $w = o$,

$$\begin{aligned}\nabla_{\boldsymbol{u}_w} \boldsymbol{J} = \nabla_{\boldsymbol{u}_o} \boldsymbol{J} &= -\boldsymbol{v}_c + \frac{\exp(\boldsymbol{u}_o^T \boldsymbol{v}_c) \cdot \boldsymbol{v}_c}{\sum_{w' \in Vocab} \exp(\boldsymbol{u}_{w'}^T \boldsymbol{v}_c)} \\ &= -\boldsymbol{v}_c + \hat{y}_o \cdot \boldsymbol{v}_c \\ &= \boldsymbol{v}_c(\hat{y}_o - 1) \\ &= \boldsymbol{v}_c(\hat{y}_w - 1)\,.\end{aligned}$$

When $w \neq o$,

$$\begin{aligned}\nabla_{\boldsymbol{u}_w} \boldsymbol{J} &= \boldsymbol{0} + \frac{\exp(\boldsymbol{u}_w^T \boldsymbol{v}_c) \cdot \boldsymbol{v}_c}{\sum_{w' \in Vocab} \exp(\boldsymbol{u}_{w'}^T \boldsymbol{v}_c)} \\ &= \boldsymbol{v}_c \cdot \hat{y}_w\,.\end{aligned}$$

Combining the two cases, we have

$$\begin{aligned}\nabla_{\boldsymbol{u}_w} \boldsymbol{J} &= \boldsymbol{v}_c(\hat{y}_w - 1\{w = o\}) \\ &= \boldsymbol{v}_c(\hat{y}_w - y_w)\,,\end{aligned}$$

for all w. Hence,

$$\nabla_{\boldsymbol{U}} \boldsymbol{J} = \boldsymbol{v}_c(\hat{\boldsymbol{y}} - \boldsymbol{y})^T\,.$$

Problem 1(d). Consider the case when $\boldsymbol{x} \in \mathbb{R}$ is a scalar, we have

$$\begin{aligned}\sigma(\boldsymbol{x}) &= \frac{1}{1 + e^{-\boldsymbol{x}}} \\ \Longrightarrow \sigma'(\boldsymbol{x}) &= \frac{-1}{(1 + e^{-\boldsymbol{x}})^2}(-e^{-\boldsymbol{x}}) \\ &= \frac{e^{-\boldsymbol{x}} + 1 - 1}{(1 + e^{-\boldsymbol{x}})^2} \\ &= \left(\frac{1}{1 + e^{-\boldsymbol{x}}}\right)\left(1 - \frac{1}{1 + e^{-\boldsymbol{x}}}\right) \\ &= \sigma(\boldsymbol{x})\big(1 - \sigma(\boldsymbol{x})\big)\,.\end{aligned}$$

When $\boldsymbol{x}$ is a vector, the sigmoid function is simply applied to it elementwise, so

$$\nabla_{\boldsymbol{x}} \sigma(\boldsymbol{x}) = diag\Big(\sigma(\boldsymbol{x}) \odot \big(1 - \sigma(\boldsymbol{x})\big)\Big)\,,$$

where $\odot$ is the elementwise product.

Problem 1(e).

$$
\begin{aligned}
\boldsymbol{J} &= \boldsymbol{J}_{neg-sample}(\boldsymbol{v}_c, o, \boldsymbol{U}) \\
&= -\log\big(\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)\big) - \sum_{k=1}^{K}\log\big(\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c)\big)\,.
\end{aligned}
$$

In the following, we will use the identity $\sigma(x) + \sigma(-x) = 1$ to simplify the expressions. Firstly,

$$
\begin{aligned}
&\nabla_{\boldsymbol{v}_c}\boldsymbol{J} \\
&= -\frac{1}{\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)}\cdot\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)\cdot\big(1-\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)\big)\cdot\boldsymbol{u}_o - \sum_{k=1}^{K}\frac{1}{\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c)}\cdot\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c)\cdot\big(1-\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c)\big)(-\boldsymbol{u}_k) \\
&= -\big(1-\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)\big)\cdot\boldsymbol{u}_o + \sum_{k=1}^{K}\big(1-\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c)\big)\cdot\boldsymbol{u}_k \\
&= -\sigma(-\boldsymbol{u}_o^T\boldsymbol{v}_c)\cdot\boldsymbol{u}_o + \sum_{k=1}^{K}\sigma(\boldsymbol{u}_k^T\boldsymbol{v}_c)\cdot\boldsymbol{u}_k\,.
\end{aligned}
$$

Secondly,

$$
o \notin \{w_1, \cdots, w_K\} \implies \boldsymbol{u}_o \neq \boldsymbol{u}_k \quad \forall k = 1, \cdots, K\,,
$$

so we have

$$
\begin{aligned}
\nabla_{\boldsymbol{u}_o}\boldsymbol{J} &= -\frac{1}{\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)}\cdot\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)\cdot\big(1-\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)\big)\cdot\boldsymbol{v}_c - \boldsymbol{0} \\
&= -\sigma(-\boldsymbol{u}_o^T\boldsymbol{v}_c)\cdot\boldsymbol{v}_c\,.
\end{aligned}
$$

Lastly, consider the multiset $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_K\}$, where the element $\boldsymbol{u}_k$ has multiplicity n_k. Therefore,

$$
\begin{aligned}
\nabla_{\boldsymbol{u}_k}\boldsymbol{J} &= \boldsymbol{0} - \sum_{\substack{1\le k'\le K \\ \boldsymbol{u}_{k'}=\boldsymbol{u}_k}}\frac{1}{\sigma(-\boldsymbol{u}_{k'}^T\boldsymbol{v}_c)}\cdot\sigma(-\boldsymbol{u}_{k'}^T\boldsymbol{v}_c)\cdot\big(1-\sigma(-\boldsymbol{u}_{k'}^T\boldsymbol{v}_c)\big)(-\boldsymbol{v}_c) \\
&= n_k\cdot\sigma(\boldsymbol{u}_k^T\boldsymbol{v}_c)\cdot\boldsymbol{v}_c\,.
\end{aligned}
$$

Gradient computation in naive-softmax requires summing over $\mathcal{O}(|Vocab|)$ terms, whereas that of negative sampling requires $\mathcal{O}(K)$ terms only, where $K << |Vocab|$.

Problem 1(f)(i).

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

Problem 1(f)(ii).

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

Problem 1(f)(iii).

$$\frac{\partial \boldsymbol{J}_{skip-gram}}{\partial \boldsymbol{v}_w} = \boldsymbol{0},$$

when $w \neq c$, since $\boldsymbol{v}_c$ is the only column vector in $\boldsymbol{V}$ related to $\boldsymbol{J}_{skip-gram}$.

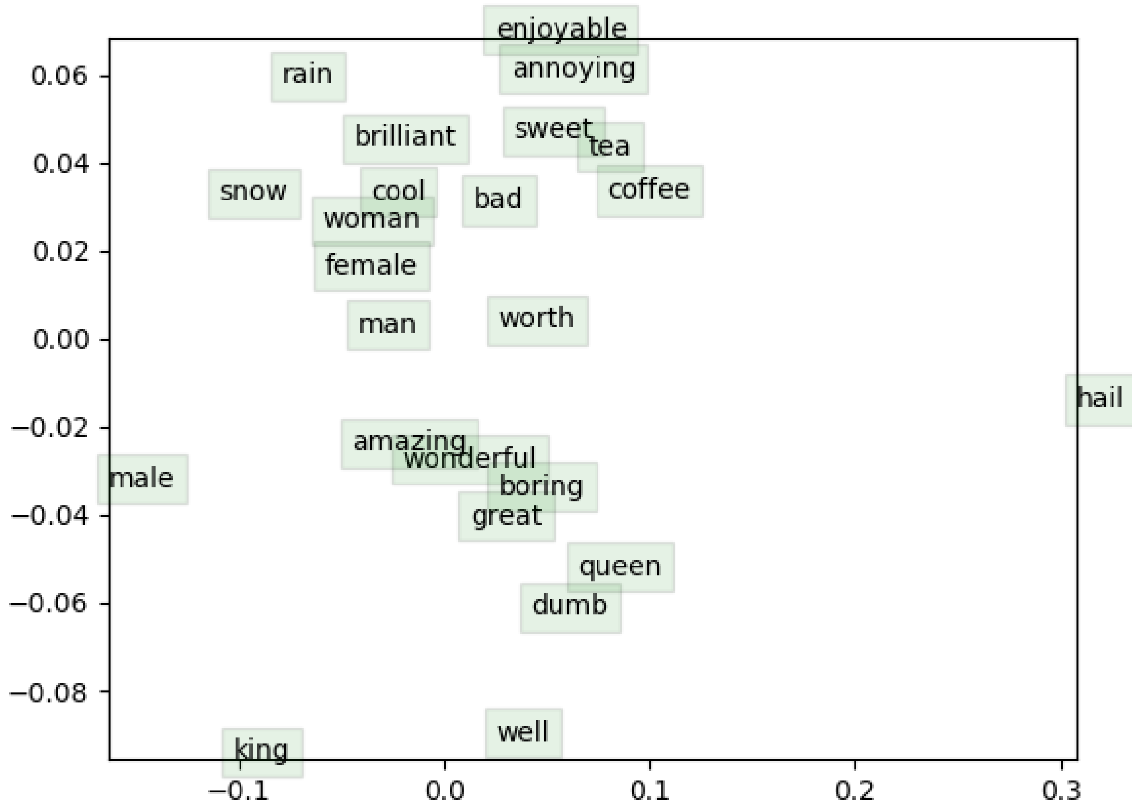

Figure 1: Visualization of Word Vectors

Problem 2(c). Some synonyms cluster together, such as {"amazing", "wonderful", "great"} and {"woman", "female"}. However, antonyms may also be close to such clusters, for example, "boring" is next to {"amazing", "wonderful", "great"} and "man" is next to {"woman", "female"}. Analogies, such as "man : king :: woman : queen", may not hold in this 2D plot.