Assignment 2 - Written: Understanding word2vec

Notation:

- $\mathbf{v}$ = input vector
- $\mathbf{u}$ = output vector
- o = The index of the desired context (outside) word.
- w = The w -th word in the vocabulary
- c = The index of the center word.

Definitions:

- The probability of an outside context word, u_o, given the center word is defined as:

$$\hat{y}_o = p(\mathbf{u_o}|\mathbf{v_c}) = \frac{e^{\mathbf{u_o}^T \mathbf{v_c}}}{\sum_{w \in \text{Vocab}} e^{\mathbf{u_w}^T \mathbf{v_c}}}$$

(a) Show that naive-softmax loss is the same as the cross-entropy loss between y and $\hat{y}$, i.e. show that:

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

Solution: B/c y is a one-hot encoded vector with zeros everywhere except at the index $w = o$ where $y_o = 1$, the sum on the LHS breaks down as follows:

$$-\sum_{w \in Vocab} y_w log(\hat{y}_w) = -(y_1 log(\hat{y}_1) + ... + y_o log(\hat{y}_o) + ... + y_{|V|} log(\hat{y}_{|V|})) = -log(\hat{y}_o)$$

(b) Compute the partial derivative of $J_{Naive-Softmax}(\mathbf{v_c}, o, U)$ w.r.t. $\mathbf{v_c}$:
Solution:

$$\begin{aligned}
\frac{\partial J_{Naive-Softmax}}{\partial \mathbf{v_c}} &= \frac{\partial}{\partial \mathbf{v_c}}[-log(\hat{y}_o)] \\
&= \frac{\partial}{\partial \mathbf{v_c}}[-log(\frac{e^{\mathbf{u_o}^T \mathbf{v_c}}}{\sum_{w \in Vocab} e^{\mathbf{u_w}^T \mathbf{v_c}}})] \\
&= -\frac{\partial}{\partial \mathbf{v_c}}[log(e^{\mathbf{u_o}^T \mathbf{v_c}}) - log(\sum_{w \in Vocab} e^{\mathbf{u_w}^T \mathbf{v_c}})] \\
&= -\frac{\partial}{\partial \mathbf{v_c}}[log(e^{\mathbf{u_o}^T \mathbf{v_c}})] + \frac{\partial}{\partial \mathbf{v_c}}[log(\sum_{w \in Vocab} e^{\mathbf{u_w}^T \mathbf{v_c}})] \\
&= -\frac{\partial}{\partial \mathbf{v_c}}[\mathbf{u_o}^T \mathbf{v_c}] + \frac{\partial}{\partial \mathbf{v_c}}[log(\sum_{w \in Vocab} e^{\mathbf{u_w}^T \mathbf{v_c}})] \\
&= -(\mathbf{u_o}) + (\frac{1}{\sum_{w \in Vocab} e^{\mathbf{u_w}^T \mathbf{v_c}}} \sum_{x \in Vocab} \mathbf{u_x} \cdot e^{\mathbf{u_x}^T \mathbf{v_c}}) \\
&= -\mathbf{u_o} + \sum_{x \in Vocab} \frac{e^{\mathbf{u_x}^T \mathbf{v_c}}}{\sum_{w \in Vocab} e^{\mathbf{u_w}^T \mathbf{v_c}}} \mathbf{u_x} \\
&= -\mathbf{u_o} + \sum_{x \in Vocab} p(\mathbf{u_x}|\mathbf{v_c}) \mathbf{u_x} \\
&= -\mathbf{u_o} + \sum_{x \in Vocab} \hat{y}_x \mathbf{u_x}
\end{aligned}$$

This says that the slope of the loss function w.r.t. the center word is equal to the difference between the observed representation of the outside context word and the expected context word according to our model.

(c) Compute the partial derivatives of $J_{naive-softmax}(\mathbf{v_c}, o, U)$ w.r.t. each of the 'outside' word vectors, $\mathbf{u_w}$'s. There will be two cases: when $w = o$, the true 'outside' word vector, and $w \neq o$, for all other words. Please write your answer in terms of y, $\hat{y}$, and $\mathbf{v_c}$.

case 1 - the outside word vector is the true context word vector:

$$\begin{aligned}
\frac{\partial J_{naive-softmax}}{\partial u_{w=o}} &= \frac{\partial}{\partial u_{w=o}}[-log(\hat{y}_o)] \\
&= \frac{\partial}{\partial u_{w=o}}[-log(\frac{e^{\mathbf{u_o}^T\mathbf{v_c}}}{\sum_{w\in Vocab} e^{\mathbf{u_w}^T\mathbf{v_c}}})] \\
&= -\frac{\partial}{\partial u_{w=o}}[\mathbf{u_o}^T\mathbf{v_c}] + \frac{\partial}{\partial u_{w=o}}[log(\sum_{w\in Vocab} e^{\mathbf{u_w}^T\mathbf{v_c}})] \\
&= -(\mathbf{v_c}) + (\frac{1}{\sum_{w\in Vocab} e^{\mathbf{u_w}^T\mathbf{v_c}}}(e^{\mathbf{u_o}^T\mathbf{v_c}} \cdot \mathbf{v_c})) \\
&= \mathbf{v_c}(\hat{y_o} - 1)
\end{aligned}$$

case 2 - the outside word vector is any context word but the true one:

$$\begin{aligned}
\frac{\partial J_{naive-softmax}}{\partial \mathbf{u_{w\neq o}}} &= \frac{\partial}{\partial \mathbf{u_{w\neq o}}}[-log(\hat{y}_o)] \\
&= \frac{\partial}{\partial \mathbf{u_{w\neq o}}}[-log(\frac{e^{\mathbf{u_o}^T\mathbf{v_c}}}{\sum_{w\in Vocab} e^{\mathbf{u_w}^T\mathbf{v_c}}})] \\
&= -\frac{\partial}{\partial \mathbf{u_{w\neq o}}}[\mathbf{u_o}^T\mathbf{v_c}] + \frac{\partial}{\partial \mathbf{u_{w\neq o}}}[log(\sum_{w\in Vocab} e^{\mathbf{u_w}^T\mathbf{v_c}})] \\
&= 0 + (\frac{1}{\sum_{w\in Vocab} e^{\mathbf{u_w}^T\mathbf{v_c}}} \cdot e^{u_{w\neq o}^T\mathbf{v_c}} \cdot \mathbf{v_c}) \\
&= \mathbf{v_c} \cdot \hat{y}_{w\neq o}
\end{aligned}$$

(**d)** Compute the derivative of the sigmoid $\sigma(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}}} = \frac{e^{\mathbf{x}}}{e^{\mathbf{x}}+1}$ w.r.t. $\mathbf{x}$, where $\mathbf{x}$ is a vector.

Solution:

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

$$J_{neg-sample}(\mathbf{v_c}, o, U) = -log(\sigma(\mathbf{u_o}^T\mathbf{v_c})) - \sum_{k=1}^{K} log(\sigma(-\mathbf{u_k}^T\mathbf{v_c}))$$

(e) Repeat parts **b** and **c**, computing partial derivatives of $J_{neg-sample}$ w.r.t. to $\mathbf{v_c}$, w.r.t. $\mathbf{u_o}$ and w.r.t. $\mathbf{u_k}$. Write your answer in terms of the vectors, $\mathbf{u_o}$, $\mathbf{v_c}$, and $\mathbf{u_k}$, where $k \in [1, K]$. After you've done this, describe why this function is much more efficient to compute than the naive-softmax loss.

(1) **w.r.t $\mathbf{v_c}$: The embedded word vector of the center word.**

$$\begin{aligned}\frac{\partial J_{neg-sample}}{\partial \mathbf{v_c}} &= \frac{\partial}{\partial \mathbf{v_c}}[-log(\sigma(\mathbf{u_o}^T\mathbf{v_c})) - \sum_{k=1}^{K} log(\sigma(-\mathbf{u_k}^T\mathbf{v_c}))] \\ &= \frac{\partial}{\partial \mathbf{v_c}}[-log(\sigma(\mathbf{u_o}^T\mathbf{v_c}))] - \frac{\partial}{\partial \mathbf{v_c}}[\sum_{k=1}^{K} log(\sigma(-\mathbf{u_k}^T\mathbf{v_c}))]\end{aligned}$$

Derivative of first term - the log of the probability that the center word and true outside word came from the corpus data: Let $\beta = \sigma(\mathbf{u_o}^T\mathbf{v_c})$

$$\begin{aligned}&(\frac{\partial}{\partial \mathbf{v_c}}[-log(\beta)])(\frac{\partial}{\partial \mathbf{v_c}}[\beta]) \\ &= (-\frac{1}{\sigma(\mathbf{u_o}^T\mathbf{v_c})})(\mathbf{u_o}\sigma(\mathbf{u_o}^T\mathbf{v_c})(1 - \sigma(\mathbf{u_o}^T\mathbf{v_c}))) \\ &= -\mathbf{u_o}(1 - \sigma(\mathbf{u_o}^T\mathbf{v_c}))\end{aligned}$$

Derivative of second term – the sum of logs of the probabilities that the center word and outside context words did not come from the corpus data.: let $\beta = \sigma(-\mathbf{u_k}^T\mathbf{v_c})$

$$\begin{aligned}&\sum_{k=1}^{K} \frac{\partial}{\partial \mathbf{v_c}}[log(\beta)]\frac{\partial}{\partial \mathbf{v_c}}[\beta] \\ &= \sum_{k=1}^{K} \frac{-\mathbf{u_k}\sigma(-\mathbf{u_k}^T\mathbf{v_c})(1 - \sigma(-\mathbf{u_k}^T\mathbf{v_c}))}{\sigma(-\mathbf{u_k}^T\mathbf{v_c})} \\ &= -\sum_{k=1}^{K} \mathbf{u_k}(1 - \sigma(-\mathbf{u_k}^T\mathbf{v_c}))\end{aligned}$$

Final:

$$\frac{\partial J_{neg-sample}}{\partial \mathbf{v_c}} = -\mathbf{u_o}(1 - \sigma(\mathbf{u_o}^T\mathbf{v_c})) + \sum_{k=1}^{K} \mathbf{u_k}(1 - \sigma(-\mathbf{u_k}^T\mathbf{v_c}))$$

(2) **w.r.t $\mathbf{u_o}$: the embedded word vector of the outside word.**

$$\begin{aligned}
\frac{\partial J_{neg-sample}}{\partial \mathbf{u_o}} &= \frac{\partial}{\partial \mathbf{u_o}}[-log(\sigma(\mathbf{u_o}^T\mathbf{v_c})) - \sum_{k=1}^{K} log(\sigma(-\mathbf{u_k}^T\mathbf{v_c}))] \\
&= \frac{\partial}{\partial \mathbf{u_o}}[-log(\sigma(\mathbf{u_o}^T\mathbf{v_c}))] - \frac{\partial}{\partial \mathbf{u_o}}[\sum_{k=1}^{K} log(\sigma(-\mathbf{u_k}^T\mathbf{v_c}))] \\
&= \frac{\partial}{\partial \mathbf{u_o}}[-log(\sigma(\mathbf{u_o}^T\mathbf{v_c}))] - 0 \\
&= -[\frac{1}{\sigma(\mathbf{u_o}^T\mathbf{v_c})}][\sigma(\mathbf{u_o}^T\mathbf{v_c})(1-\sigma(\mathbf{u_o}^T\mathbf{v_c}))\mathbf{v_c}] \\
&= -\mathbf{v_c}(1-\sigma(\mathbf{u_o}^T\mathbf{v_c}))
\end{aligned}$$

(3) **w.r.t $\mathbf{u_k}$: the embedded word vector of one of the K negative samples.**

$$\begin{aligned}
\frac{\partial J_{neg-sample}}{\partial \mathbf{u_k}} &= \frac{\partial}{\partial \mathbf{u_k}}[-log(\sigma(\mathbf{u_o}^T\mathbf{v_c})) - \sum_{k=1}^{K} log(\sigma(-\mathbf{u_k}^T\mathbf{v_c}))] \\
&= \frac{\partial}{\partial \mathbf{u_k}}[-log(\sigma(\mathbf{u_o}^T\mathbf{v_c}))] - \frac{\partial}{\partial \mathbf{u_k}}[\sum_{k=1}^{K} log(\sigma(-\mathbf{u_k}^T\mathbf{v_c}))] \\
&= 0 - \frac{\partial}{\partial \mathbf{u_k}}[\sum_{k=1}^{K} log(\sigma(-\mathbf{u_k}^T\mathbf{v_c}))] \\
&= -\frac{-\mathbf{v_c}\sigma(-\mathbf{u_k}^T\mathbf{v_c})(1-\sigma(-\mathbf{u_k}^T\mathbf{v_c}))}{\sigma(-\mathbf{u_k}^T\mathbf{v_c})} \\
&= \mathbf{v_c}(1-\sigma(-\mathbf{u_k}^T\mathbf{v_c}))
\end{aligned}$$

Follow-up: This loss function is much more efficient to compute than the naive-softmax loss because it takes into account just K more sample word vectors ($O(K)$) whereas in the naive-softmax loss we must normalize the unnormalized probabilities, requiring that we look at all the word vectors in the entire vocabulary ($O(|V|)$).

(f) Suppose the center word is $c = w_t$ and the context window is $[w_{t-m}, ..., w_{t-1}, w_t, w_{t+1}, ..., w_{t+m}]$, where m is the context windows size. In the skip-gram version of word2vec, the total loss for the context window is:

$$J_{skip-gram}(\mathbf{v_c}, w_{t-m}, ..., w_{t+m}, \mathbf{U}) = \sum_{-m \leq j \leq m, j \neq 0} J(\mathbf{v_c}, w_{t+j}, \mathbf{U})$$

Here, $J(\mathbf{v_c}, w_{t+j}, \mathbf{U})$ represents an arbitrary loss term for the center word $c = w_t$ and outside word w_{t+j}. Write down three partials

(i) $\partial J_{skip-gram}(\mathbf{v_c}, w_{t-m}, ..., w_{t+m}, \mathbf{U})/\partial \mathbf{U}$

$$\sum_{-m \leq j \leq m, j \neq 0} \partial J_{skip-gram}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{U}$$

(ii) $\partial J_{skip-gram}(\mathbf{v_c}, w_{t-m}, ..., w_{t+m}, \mathbf{U})/\partial \mathbf{v}_c$

$$\sum_{-m \leq j \leq m, j \neq 0} \partial J_{skip-gram}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{v_c}$$

(iii) $\partial J_{skip-gram}(\mathbf{v_c}, w_{t-m}, ..., w_{t+m}, \mathbf{U})/\partial \mathbf{v}_w$ when $w \neq c$

$$\mathbf{0}$$

Code Results

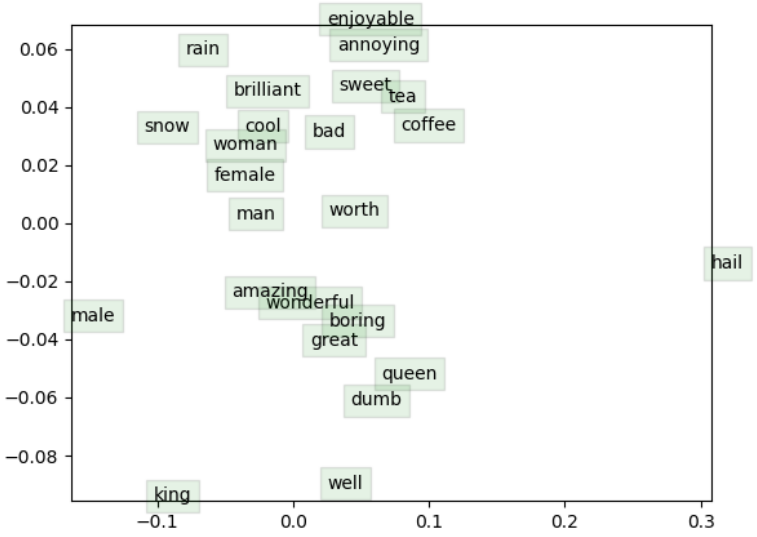

Figure 1: The learned word vectors in 2D space.