CS 224n Assignment #2: word2vec (49 Points)

Anthony Weng

Due on Tuesday Jan. 24, 2023 by **4:30pm (before class)**

1 Written: Understanding word2vec (31 points)

Recall that the key insight behind `word2vec` is that *'a word is known by the company it keeps'*. Concretely, consider a 'center' word c surrounded before and after by a context of a certain length. We term words in this contextual window 'outside words' (O). For example, in Figure 1, the context window length is 2, the center word c is 'banking', and the outside words are 'turning', 'into', 'crises', and 'as':

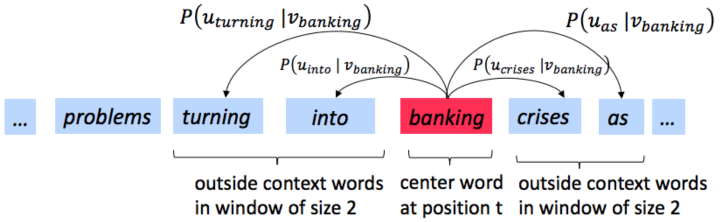
$P(u_{turning} \mid v_{banking})$
$P(u_{into} \mid v_{banking})$
$P(u_{crises} \mid v_{banking})$
$P(u_{as} \mid v_{banking})$
...
problems
turning
into
banking
crises
as
...
outside context words in window of size 2
center word at position t
outside context words in window of size 2

Figure 1: The word2vec skip-gram prediction model with window size 2

Skip-gram `word2vec` aims to learn the probability distribution $P(O|C)$. Specifically, given a specific word o and a specific word c, we want to predict $P(O = o|C = c)$: the probability that word o is an 'outside' word for c (i.e., that it falls within the contextual window of c). We model this probability by taking the softmax function over a series of vector dot-products:

$$P(O = o \mid C = c) = \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \tag{1}$$

For each word, we learn vectors u and v, where $\boldsymbol{u}_o$ is the 'outside' vector representing outside word o, and $\boldsymbol{v}_c$ is the 'center' vector representing center word c. We store these parameters in two matrices, $\boldsymbol{U}$ and $\boldsymbol{V}$. The columns of $\boldsymbol{U}$ are all the 'outside' vectors $\boldsymbol{u}_w$; the columns of $\boldsymbol{V}$ are all of the 'center' vectors $\boldsymbol{v}_w$. Both $\boldsymbol{U}$ and $\boldsymbol{V}$ contain a vector for every $w \in$ Vocabulary.[1]

Recall from lectures that, for a single pair of words c and o, the loss is given by:

$$\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log P(O = o|C = c). \tag{2}$$

We can view this loss as the cross-entropy[2] between the true distribution $\boldsymbol{y}$ and the predicted distribution $\hat{\boldsymbol{y}}$, for a particular center word c and a particular outside word o. Here, both $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are vectors with length equal to the number of words in the vocabulary. Furthermore, the k^{th} entry in these vectors indicates the conditional probability of the k^{th} word being an 'outside word' for the given c. The true empirical distribution $\boldsymbol{y}$ is a one-hot vector with a 1 for the true outside word o, and 0 everywhere else, for this particular example of center word c and outside word o.[3] The predicted distribution $\hat{\boldsymbol{y}}$ is the probability distribution $P(O|C = c)$ given by our model in equation (1).

Note: Throughout this homework, when computing derivatives, please use the method reviewed during the lecture (i.e. no Taylor Series Approximations).

[1] Assume that every word in our vocabulary is matched to an integer number k. Bolded lowercase letters represent vectors. $\boldsymbol{u}_k$ is both the k^{th} column of $\boldsymbol{U}$ and the 'outside' word vector for the word indexed by k. $\boldsymbol{v}_k$ is both the k^{th} column of $\boldsymbol{V}$ and the 'center' word vector for the word indexed by k. **In order to simplify notation we shall interchangeably use k to refer to word k and the index of word k.**

[2] The **cross-entropy loss** between the true (discrete) probability distribution p and another distribution q is $-\sum_i p_i \log(q_i)$.

[3] Note that the true conditional probability distribution of context words for the entire training dataset would not be one-hot.

(a) (2 points) Prove that the naive-softmax loss (Equation 2) is the same as the cross-entropy loss between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, i.e. (note that $\boldsymbol{y}, \hat{\boldsymbol{y}}$ are vectors and $\hat{\boldsymbol{y}}_o$ is a scalar):

$$-\sum_{w \in \text{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) = -\log(\hat{\boldsymbol{y}}_o). \tag{3}$$

Your answer should be one line. You may describe your answer in words.

Answer:

(a) As $\boldsymbol{y}$ is a 1-hot vector with $\boldsymbol{y}_w = 1$ when $w = o$, the only non-zero term in the cross-entropy summation is $\boldsymbol{y}_o * \log(\hat{\boldsymbol{y}_o})$ – meaning the LHS of Equ.(3) equals $-1 * \log(\hat{\boldsymbol{y}_o}) = -\log(\hat{\boldsymbol{y}_o})$ where $\hat{\boldsymbol{y}_o} = P(O = o|c)$, which shows that Equ.(3) and Equ.(2) are equivalent, as desired ■.

(b) (7 points)

(i) Compute the partial derivative of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to $\boldsymbol{v}_c$. Please write your answer in terms of $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, $\boldsymbol{U}$, and show your work to receive full credit.

- **Note**: Your final answers for the partial derivative should follow the shape convention: the partial derivative of any function $f(x)$ with respect to x should have the **same shape** as x.[4]
- Please provide your answers for the partial derivative in vectorized form. For example, when we ask you to write your answers in terms of $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, and $\boldsymbol{U}$, you may not refer to specific elements of these terms in your final answer (such as $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, $\ldots$).

(ii) When is the gradient you computed equal to zero?
Hint: You may wish to review and use some introductory linear algebra concepts.

(iii) The gradient you found is the difference between two terms. Provide an interpretation of how each of these terms improves the word vector when this gradient is subtracted from the word vector v_c.

(iv) In many downstream applications using word embeddings, L2 normalized vectors (e.g. $\mathbf{u}/||\mathbf{u}||_2$ where $||\mathbf{u}||_2 = \sqrt{\sum_i u_i^2}$) are used instead of their raw forms (e.g. $\mathbf{u}$). Now, suppose you would like to classify phrases as being positive or negative. When would L2 normalization take away useful information for the downstream task? When would it not? Hint: Consider the case where $\mathbf{u}_x = \alpha \mathbf{u}_y$ for some words $x \neq y$ and some scalar α.

[4] This allows us to efficiently minimize a function using gradient descent without worrying about reshaping or dimension mismatching. While following the shape convention, we're guaranteed that $\theta := \theta - \alpha \frac{\partial J(\theta)}{\partial \theta}$ is a well-defined update rule.

Answer:

(b)(i) First, we re-express $\boldsymbol{J}_{\textsf{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ as follows:

$$\begin{aligned}
\boldsymbol{J}_{\textsf{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= -\log P(O = o|C = c) \\
&= -\sum_{w\in\textsf{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) = -\log(\hat{\boldsymbol{y}}_o) \\
&= -\log(\frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w\in\textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}) \\
&= -\log(\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) + \log(\sum_{w\in\textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)) \\
&= -(\boldsymbol{u}_o^\top \boldsymbol{v}_c) + \log(\sum_{w\in\textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c))
\end{aligned}$$

Then, we evaluate $\frac{\partial \boldsymbol{J}_{\textsf{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c}$ as follows:

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\textsf{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} &= -\frac{\partial}{\partial \boldsymbol{v}_c}(\boldsymbol{u}_o^\top \boldsymbol{v}_c) + \frac{\partial}{\partial \boldsymbol{v}_c}\log(\sum_{w\in\textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)) \\
&= -\boldsymbol{u}_o + \frac{1}{\sum_{w'\in\textsf{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} * \sum_{w\in\textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) * \boldsymbol{u}_w \\
&= -\boldsymbol{u}_o + \sum_{w\in\textsf{Vocab}} \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) * \boldsymbol{u}_w}{\sum_{w'\in\textsf{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \\
&= -\boldsymbol{u}_o + \sum_{w\in\textsf{Vocab}} \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w'\in\textsf{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} * \boldsymbol{u}_w \\
&= -\boldsymbol{u}_o + \sum_{w\in\textsf{Vocab}} P(O = w \mid C = c) * \boldsymbol{u}_w = -\boldsymbol{u}_o + \sum_{w\in\textsf{Vocab}} \hat{\boldsymbol{y}}_w \boldsymbol{u}_w \\
&= \sum_{w\in\textsf{Vocab}} \hat{\boldsymbol{y}}_w \boldsymbol{u}_w - \boldsymbol{u}_o = \boldsymbol{U} \times \hat{\boldsymbol{y}} - \boldsymbol{U} \times \boldsymbol{y} \\
&= \boldsymbol{U} \times (\hat{\boldsymbol{y}} - \boldsymbol{y}) \ \blacksquare.
\end{aligned}$$

(b)(ii) The gradient found in (b)(i) is equal to 0 when $\boldsymbol{U} = \boldsymbol{O}$ and/or $(\hat{\boldsymbol{y}} - \boldsymbol{y}) = \boldsymbol{0}$. Under the assumption that we learn/initialize non-zero outside vectors for each word w, then $\boldsymbol{U} \neq \boldsymbol{O}$. Thus, a gradient of zero is only achieved when $(\hat{\boldsymbol{y}} - \boldsymbol{y}) = \boldsymbol{0}$; i.e., when the predicted and true outside word distributions are identical (equivalently, when the model predicted the observed true outside word with full confidence; it makes sense that the gradient is zero here, since we wouldn't want a gradient-wise correction if the model achieved a perfect prediction).

(b)(iii) When this gradient is subtracted from the word vector v_c (i.e., $\boldsymbol{v}_c - \boldsymbol{U} \times (\hat{\boldsymbol{y}} - \boldsymbol{y}) = \boldsymbol{v}_c - \boldsymbol{U}\hat{\boldsymbol{y}} + \boldsymbol{U}y$), the center word vector v_c improves by: (1) becoming generally less similar to all of the outside word vectors, proportional to how strongly we believe some word is an outside word given the center word c (i.e., $-\boldsymbol{U}\hat{\boldsymbol{y}}$); while also (2) becoming more similar to the outside word vector for the word o which did appear in the center word's context (i.e. $+\boldsymbol{U}\boldsymbol{y}$).

(b)(iv) L2 normalization *would* take away useful information for the downstream task if we are interested in some notion of significance/consistency in addition to the positive/negative nature of phrases. Consider two positive-associated words x and y: if x and y show up in the exact same contexts in the training corpus (e.g., "The x = compassionate person is nice." and "The y = kind person is nice."), but x shows up in twice as many of these contexts (e.g., "The compassionate person is nice" shows up in the corpus twice as many times as "The kind person is nice"), then $\boldsymbol{u}_x$ will be a vector of greater magnitude than $\boldsymbol{u}_y$ but have the same direction (i.e. $\boldsymbol{u}_x = \alpha \boldsymbol{u}_y$ for some $\alpha > 1$). This relationship between $\boldsymbol{u}_x$ and $\boldsymbol{u}_y$ carries information about the degree to/consistency with which x indicates a positive phrase relative to y, but this information would be erased by the L2 normalization of $\boldsymbol{u}_x, \boldsymbol{u}_y$.

In short, L2 normalization would take away useful information when degree/consistency of sentiment is of interest in addition to the positive/negative classification of phrases, while L2 normalization would not take away useful information when our downstream application is solely interested in the directionality (i.e., positive or negative) of our phrase classifications.

(c) (5 points) Compute the partial derivatives of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to each of the 'outside' word vectors, $\boldsymbol{u}_w$'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 $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, and $\boldsymbol{v}_c$. In this subpart, you may use specific elements within these terms as well (such as $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, $\ldots$). Note that $\boldsymbol{u}_w$ is a vector while $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ are scalars. Show your work to receive full credit.

Answer:

(c) As a preface, we note $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -(\boldsymbol{u}_o^\top \boldsymbol{v}_c) + \log(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c))$ (as demonstrated in (b)(i)).

First consider the case when $w = o$:

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_w} &= -\frac{\partial}{\partial \boldsymbol{u}_w}(\boldsymbol{u}_o^\top \boldsymbol{v}_c) + \frac{\partial}{\partial u_w} \log(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)) \\
&= -\boldsymbol{v}_c + \frac{1}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} * \frac{\partial}{\partial \boldsymbol{u}_w} \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \\
&= -\boldsymbol{v}_c + \frac{1}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} * \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \boldsymbol{v}_c \\
&= -\boldsymbol{v}_c + \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} * \boldsymbol{v}_c \\
&= -\boldsymbol{v}_c + P(O = w \mid C = c) * \boldsymbol{v}_c \\
&= -\boldsymbol{v}_c + \hat{\boldsymbol{y}}_w * \boldsymbol{v}_c \\
&= \boldsymbol{v}_c(\hat{\boldsymbol{y}}_w - 1)
\end{aligned}$$

Then, consider the case when $w \neq o$:

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_w} &= -\frac{\partial}{\partial \boldsymbol{u}_w}(\boldsymbol{u}_o^\top \boldsymbol{v}_c) + \frac{\partial}{\partial u_w} \log(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)) \\
&= 0 + \frac{1}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} * \frac{\partial}{\partial \boldsymbol{u}_w} \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \\
&= \frac{1}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} * \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \boldsymbol{v}_c \\
&= \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} * \boldsymbol{v}_c \\
&= P(O = w \mid C = c) * \boldsymbol{v}_c \\
&= \hat{\boldsymbol{y}}_w * \boldsymbol{v}_c \\
&= \boldsymbol{v}_c \hat{\boldsymbol{y}}_w
\end{aligned}$$

Combining the cases:

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_w} = \begin{cases} \boldsymbol{v}_c(\hat{\boldsymbol{y}}_w - 1) & w = o \\ \boldsymbol{v}_c \hat{\boldsymbol{y}}_w & w \neq o \end{cases}$$

(d) (1 point) Write down the partial derivative of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to $\boldsymbol{U}$. Please break down your answer in terms of the column vectors $\frac{\partial \boldsymbol{J}(\boldsymbol{v}_c,o,\boldsymbol{U})}{\partial \boldsymbol{u}_1}$, $\frac{\partial \boldsymbol{J}(\boldsymbol{v}_c,o,\boldsymbol{U})}{\partial \boldsymbol{u}_2}$, $\cdots$, $\frac{\partial \boldsymbol{J}(\boldsymbol{v}_c,o,\boldsymbol{U})}{\partial \boldsymbol{u}_{|\text{Vocab}|}}$. No derivations are necessary, just an answer in the form of a matrix.

Answer:

(d)

$$\begin{aligned}\frac{\partial \boldsymbol{J}_{\textsf{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{U}} &= \left[\frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_1}, \cdots, \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_k}, \cdots, \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_{|\textsf{Vocab}|}}\right] \\ &= \left[\boldsymbol{v}_c \hat{\boldsymbol{y}}_1, \cdots, \boldsymbol{v}_c(\hat{\boldsymbol{y}}_k - 1), \cdots, \boldsymbol{v}_c \hat{\boldsymbol{y}}_{|\textsf{Vocab}|}\right]\end{aligned}$$

where word k (with outside word vector $\boldsymbol{u}_k$) satisfies $k = o$, and for other words $i \in [1, \cdots, |\text{Vocab}|]$, word $i \neq o$.

(e) (2 points) The Leaky ReLU (Leaky Rectified Linear Unit) activation function is given by Equation 4 and Figure 2:

$$f(x) = \max(\alpha x, x) \tag{4}$$

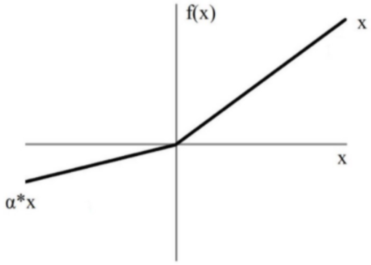

Figure 2: Leaky ReLU

Where x is a scalar and $0 < \alpha < 1$, please compute the derivative of $f(x)$ with respect to x. You may ignore the case where the derivative is not defined at 0.[5]

Answer:

(e)

$$\tfrac{d}{dx} f(x) = \begin{cases} 1 & x > 0 \\ \alpha & x < 0 \end{cases}$$

[5] If you're interested in how to handle the derivative at this point, you can read more about the notion of subderivatives.

(f) (3 points) The sigmoid function is given by Equation 5:

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \tag{5}$$

Please compute the derivative of $\sigma(x)$ with respect to x, where x is a scalar. Please write your answer in terms of $\sigma(x)$. Show your work to receive full credit.

Answer:

(f)

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

(g) (6 points) Now we shall consider the Negative Sampling loss, which is an alternative to the Naive Softmax loss. Assume that K negative samples (words) are drawn from the vocabulary. For simplicity of notation we shall refer to them as $w_1, w_2, \dots, w_K$, and their outside vectors as $\boldsymbol{u}_{w_1}, \boldsymbol{u}_{w_2}, \dots, \boldsymbol{u}_{w_K}$.[6] For this question, assume that the K negative samples are distinct. In other words, $i \neq j$ implies $w_i \neq w_j$ for $i, j \in \{1, \dots, K\}$. Note that $o \notin \{w_1, \dots, w_K\}$. For a center word c and an outside word o, the negative sampling loss function is given by:

$$\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \sum_{s=1}^{K} \log(\sigma(-\boldsymbol{u}_{w_s}^\top \boldsymbol{v}_c)) \tag{6}$$

for a sample $w_1, \dots w_K$, where $\sigma(\cdot)$ is the sigmoid function.[7]

(i) Please repeat parts (b) and (c), computing the partial derivatives of $\boldsymbol{J}_{\text{neg-sample}}$ with respect to $\boldsymbol{v}_c$, with respect to $\boldsymbol{u}_o$, and with respect to the s^{th} negative sample $\boldsymbol{u}_{w_s}$. Please write your answers in terms of the vectors $\boldsymbol{v}_c$, $\boldsymbol{u}_o$, and $\boldsymbol{u}_{w_s}$, where $s \in [1, K]$. Show your work to receive full credit. **Note:** you should be able to use your solution to part (f) to help compute the necessary gradients here.

(ii) In lecture, we learned that an efficient implementation of backpropagation leverages the re-use of previously-computed partial derivatives. Which quantity could you reuse amongst the three partial derivatives calculated above to minimize duplicate computation? Write your answer in terms of $\boldsymbol{U}_{o,\{w_1,\dots,w_K\}} = \left[\boldsymbol{u}_o, -\boldsymbol{u}_{w_1}, \dots, -\boldsymbol{u}_{w_K}\right]$, a matrix with the outside vectors stacked as columns, and $\mathbf{1}$, a $(K+1) \times 1$ vector of 1's.[8] Additional terms and functions (other than $\boldsymbol{U}_{o,\{w_1,\dots,w_K\}}$ and $\mathbf{1}$) can be used in your solution.

(iii) Describe with one sentence why this loss function is much more efficient to compute than the naive-softmax loss.

Caveat: So far we have looked at re-using quantities and approximating softmax with sampling for faster gradient descent. Do note that some of these optimizations might not be necessary on modern GPUs and are, to some extent, artifacts of the limited compute resources available at the time when these algorithms were developed.

[6] Note: In the notation for parts (g) and (h), we are using words, not word indices, as subscripts for the outside word vectors.

[7] Note: The loss function here is the negative of what Mikolov et al. had in their original paper, because we are doing a minimization instead of maximization in our assignment code. Ultimately, this is the same objective function.

[8] Note: NumPy will automatically broadcast 1 to a vector of 1's if the computation requires it, so you generally don't have to construct **1** on your own during implementation.

Answer:

(g)(i) $\frac{\boldsymbol{J}_{\textsf{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c}$:

$$= -\frac{1}{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)} * \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c) * (1 - \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) * \boldsymbol{u}_o - \sum_{s=1}^{K} \frac{1}{\sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)} * \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c) * (1 - \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)) * -\boldsymbol{u}_{w_n}^\top$$

$$= -(1 - \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) * \boldsymbol{u}_o - \sum_{s=1}^{K} (1 - \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)) * -\boldsymbol{u}_{w_n}^\top$$

$$= (\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c) - 1)\boldsymbol{u}_o - \sum_{s=1}^{K} (\sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c) - 1) \cdot \boldsymbol{u}_{w_n} \ \blacksquare.$$

$\frac{\boldsymbol{J}_{\textsf{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o}$:

$$= -\frac{1}{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)} * \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c) * (1 - \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) * \boldsymbol{v}_c - \sum_{s=1}^{K} \frac{1}{\sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)} * \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c) * (1 - \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)) * 0$$

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

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

$\frac{\boldsymbol{J}_{\textsf{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_{w_s}}$:

$$= -\frac{1}{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)} * \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c) * (1 - \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) * 0 - \sum_{s'=1}^{K} \frac{1}{\sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)} * \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c) * (1 - \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)) * -\boldsymbol{v}_c$$

$$= -\sum_{s'=1}^{K} \begin{cases} 0 \text{ if } s' \neq s \\ (1 - \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)) * -\boldsymbol{v}_c \text{ if } s' = s \end{cases}$$

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

$$= -(\sigma(-\boldsymbol{u}_{w_s}^\top \boldsymbol{v}_c) - 1)\boldsymbol{v}_c \ \blacksquare.$$

(g)(ii) To minimize duplicate computation amongst the three partial derivatives, we could compute the following quantity once:

$$\sigma(\boldsymbol{U}_{o,\{w_1,\ldots,w_K\}}{}^\top \times \boldsymbol{v}_c) - \boldsymbol{1}$$

where $\sigma(\boldsymbol{X})$ denotes the element-wise application of the sigmoid function to $\boldsymbol{X}$.

(g)(iii) For each loss computation, use of naive-softmax loss requires $O(|\textsf{Vocab}|)$ matrix operations (due to the computation of $\sum_{w \in \textsf{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)$) whereas use of negative sampling loss only requires $O(K)$ operations (and depending on choice of K, $K << |\textsf{Vocab}|$).

(h) (2 points) Now we will repeat the previous exercise, but without the assumption that the K sampled words are distinct. Assume that K negative samples (words) are drawn from the vocabulary. For simplicity of notation we shall refer to them as $w_1, w_2, \dots, w_K$ and their outside vectors as $\boldsymbol{u}_{w_1}, \dots, \boldsymbol{u}_{w_K}$. In this question, you may not assume that the words are distinct. In other words, $w_i = w_j$ may be true when $i \neq j$ is true. Note that $o \notin \{w_1, \dots, w_K\}$. For a center word c and an outside word o, the negative sampling loss function is given by:

$$\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \sum_{s=1}^{K} \log(\sigma(-\boldsymbol{u}_{w_s}^\top \boldsymbol{v}_c)) \tag{7}$$

for a sample $w_1, \dots w_K$, where $\sigma(\cdot)$ is the sigmoid function.

Compute the partial derivative of $\boldsymbol{J}_{\text{neg-sample}}$ with respect to a negative sample $\boldsymbol{u}_{w_s}$. Please write your answers in terms of the vectors $\boldsymbol{v}_c$ and $\boldsymbol{u}_{w_s}$, where $s \in [1, K]$. Show your work to receive full credit. Hint: break up the sum in the loss function into two sums: a sum over all sampled words equal to w_s and a sum over all sampled words not equal to w_s. Notation-wise, you may write 'equal' and 'not equal' conditions below the summation symbols, such as in Equation 8.

Answer:

(h) $\frac{\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_{w_s}}$:

$$= -\frac{1}{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)} * \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c) * (1 - \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) * 0 - \sum_{s'=1}^{K} \frac{1}{\sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)} * \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c) * (1 - \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)) * \frac{\partial}{\partial \boldsymbol{u}_{w_s}}(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c))$$

$$= -\left(\sum_{\substack{s'=1, \\ w_{s'} = w_s}}^{K} (1 - \sigma(-\boldsymbol{u}_{w_n}^\top \boldsymbol{v}_c)) * -\boldsymbol{v}_c + \sum_{\substack{s'=1, \\ w_{s'} \neq w_s}}^{K} 0 \right)$$

$$= -\sum_{\substack{s'=1, \\ w_{s'} = w_s}}^{K} (1 - \sigma(-\boldsymbol{u}_{w_s}^\top \boldsymbol{v}_c)) * -\boldsymbol{v}_c$$

$$= -\sum_{\substack{s'=1, \\ w_{s'} = w_s}}^{K} (\sigma(-\boldsymbol{u}_{w_s}^\top \boldsymbol{v}_c) - 1)\boldsymbol{v}_c \ \blacksquare.$$

(i) (3 points) Suppose the center word is $c = w_t$ and the context window is $[w_{t-m}, \ldots, w_{t-1}, w_t, w_{t+1}, \ldots, w_{t+m}]$, where m is the context window size. Recall that for the skip-gram version of `word2vec`, the total loss for the context window is:

$$\boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U}) = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U}) \tag{8}$$

Here, $\boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$ represents an arbitrary loss term for the center word $c = w_t$ and outside word w_{t+j}. $\boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$ could be $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$ or $\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$, depending on your implementation.

Write down three partial derivatives:

(i) $\frac{\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{U}}$

(ii) $\frac{\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_c}$

(iii) $\frac{\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_w}$ when $w \ne c$

Write your answers in terms of $\frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{U}}$ and $\frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{v}_c}$. This is very simple – each solution should be one line.

Answer:

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

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

(i)(iii) When $w \ne c$: $\frac{\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_w} = 0$

Once you're done: *Given that you computed the derivatives of $\boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$ with respect to all the model parameters $\boldsymbol{U}$ and $\boldsymbol{V}$ in parts (a) to (c), you have now computed the derivatives of the full loss function $\boldsymbol{J}_{\textit{skip-gram}}$ with respect to all parameters. You're ready to implement `word2vec`!*

2 Coding: Implementing word2vec (18 points)

In this part you will implement the word2vec model and train your own word vectors with stochastic gradient descent (SGD). Before you begin, first run the following commands within the assignment directory in order to create the appropriate conda virtual environment. This guarantees that you have all the necessary packages to complete the assignment. **Windows users** may wish to install the Linux Windows Subsystem[9]. Also note that you probably want to finish the previous math section before writing the code since you will be asked to implement the math functions in Python. You'll probably want to implement and test each part of this section in order, since the questions are cumulative.

```
conda env create -f env.yml
conda activate a2
```

Once you are done with the assignment you can deactivate this environment by running:

```
conda deactivate
```

For each of the methods you need to implement, we included approximately how many lines of code our solution has in the code comments. These numbers are included to guide you. You don't have to stick to them, you can write shorter or longer code as you wish. If you think your implementation is significantly longer than ours, it is a signal that there are some `numpy` methods you could utilize to make your code both shorter and faster. `for` loops in Python take a long time to complete when used over large arrays, so we expect you to utilize `numpy` methods. We will be checking the efficiency of your code. You will be able to see the results of the autograder when you submit your code to `Gradescope`, we recommend submitting early and often.
Note: If you are using Windows and have trouble running the .sh scripts used in this part, we recommend trying Gow or manually running commands in the scripts.

(a) (12 points) We will start by implementing methods in `word2vec.py`. You can test a particular method by running `python word2vec.py m` where `m` is the method you would like to test. For example, you can test the sigmoid method by running `python word2vec.py sigmoid`.

 (i) Implement the `sigmoid` method, which takes in a vector and applies the sigmoid function to it.
 (ii) Implement the softmax loss and gradient in the `naiveSoftmaxLossAndGradient` method.
 (iii) Implement the negative sampling loss and gradient in the `negSamplingLossAndGradient` method.
 (iv) Implement the skip-gram model in the `skipgram` method.

 When you are done, test your entire implementation by running `python word2vec.py`.

(b) (4 points) Complete the implementation for your SGD optimizer in the `sgd` method of `sgd.py`. Test your implementation by running `python sgd.py`.

(c) (2 points) Show time! Now we are going to load some real data and train word vectors with everything you just implemented! We are going to use the Stanford Sentiment Treebank (SST) dataset to train word vectors, and later apply them to a simple sentiment analysis task. You will need to fetch the datasets first. To do this, run `sh get_datasets.sh`. There is no additional code to write for this part; just run `python run.py`.

[9] https://techcommunity.microsoft.com/t5/windows-11/how-to-install-the-linux-windows-subsystem-in-windows-11/m-p/2701207

Note: The training process may take a long time depending on the efficiency of your implementation and the compute power of your machine ***(an efficient implementation takes one to two hours)****. Plan accordingly!*

After 40,000 iterations, the script will finish and a visualization for your word vectors will appear. It will also be saved as `word_vectors.png` in your project directory. **Include the plot in your homework write up.** In at most three sentences, briefly explain what you see in the plot. This may include, but is not limited to, observations on clusters and words that you expect to cluster but do not.

Answer:

Plot:

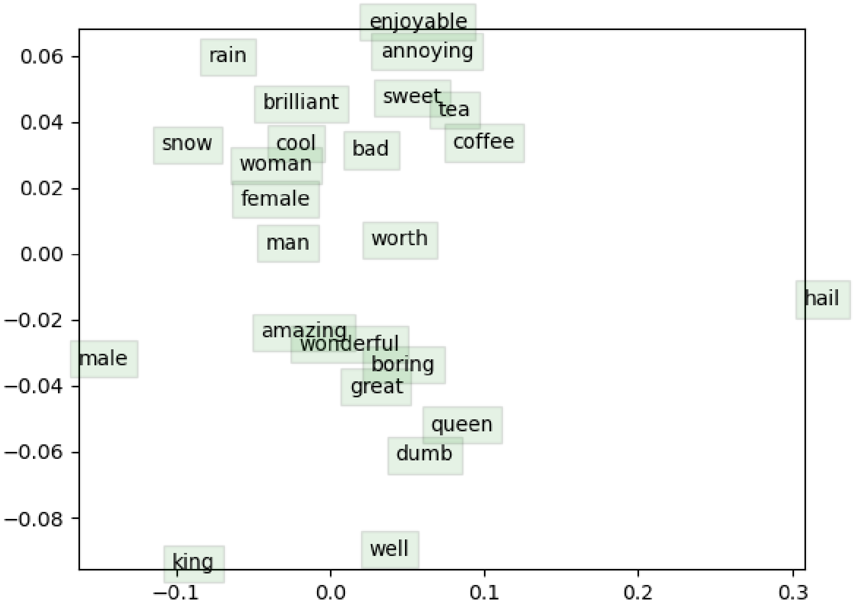

Figure 3: 2-D reduction of my `word2vec` implementation's generated vectors.

Answer:

Plot observations/summary:

1. Overall, my `word2vec` implementation appears to have generated some semi-reasonable word vectors, as we see related as well as synonymous/antonymous words (as noted in Assignment 1, antonymous words can be closer than synonymous ones!) appearing in relative proximity, meaning they were found in similar contexts, as we would expect - examples include 'tea' and 'coffee' (both beverages), 'woman' and 'female' (both words describing a gender), and 'enjoyable' and 'annoying' (antonyms which are both often used to describe an activity).
2. However, there are some groupings of word vectors which we would expect to be clustered but are not - for instance, the vector for 'hail' is far removed from those of 'rain' and 'snow,' perhaps indicative of a corpus-specific lack of content for certain words like 'hail' (and therefore, relatively inaccurate word vectors) relative to related others like 'rain' and 'snow'.
3. Finally, there are some plot features which are illustrative of potential bias in the corpus, such as the proximity of the 'queen' and 'dumb' vectors and the comparative distance between the 'king' and 'dumb' vectors - although this may partially be a product of the word vector dimensionality reduction, it could also be illustrative as to how 'queen's are described more negatively than their 'king' counterparts.

3 Submission Instructions

You shall submit this assignment on Gradescope as two submissions – one for "Assignment 2 [coding]" and another for 'Assignment 2 [written]":

(a) Run the `collect_submission.sh` script to produce your `assignment2.zip` file.

(b) Upload your `assignment2.zip` file to Gradescope to "Assignment 2 [coding]".

(c) Upload your written solutions to Gradescope to "Assignment 2 [written]".