CS224n Assignment 2: Understanding word2vec

David Lee

July 16, 2019

The given equations:

- the softmax function

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

- the loss function

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

- the sigmoid function

$$\sigma(\boldsymbol{x})=\frac{1}{1+e^{-\boldsymbol{x}}}=\frac{e^{\boldsymbol{x}}}{e^{\boldsymbol{x}}+1} \tag{3}$$

(a) Show that the naive-softmax loss given in Equeation (2) is the same as the cross-entropy loss between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$; i.e., show that

$$-\sum_{w\in\text{ Vocab}}y_w\log\left(\hat{y}_w\right)=-\log\left(\hat{y}_o\right) \tag{4}$$

Because the true label y_w is a one-hot vector. And it is 1 when w is that vocabulary, otherwise it is 0.

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

That is, the term can be reduce and represent with the negative log of its predict output vector.

(b) Compute the partial derivative of $\boldsymbol{J}_{\text{naive-softmax}}(v_c, o, U)$ with respect to $\boldsymbol{v}_c$. Please write your answer in terms of $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, and $\boldsymbol{U}$ reference 1[1] : Classification and Loss Evaluation - Softmax and Cross Entropy Loss

reference 2[2] : Derivatives of log and exp

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

(b) Compute the partial derivatives of $\boldsymbol{J}_{\text{naive-softmax}}(v_c, o, U)$ 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 you answer in terms of $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, and $\boldsymbol{v}_c$
Basically shared the same derivation of the first part thus skip some steps (partial derivative on the log)

[1] `https://deepnotes.io/softmax-crossentropy`

[2] `https://www.themathpage.com/aCalc/exponential.htm`

$$\frac{\partial \boldsymbol{J}_{\text{ naive-softmax }}(v_c, o, U)}{\partial \boldsymbol{v}_w} = \frac{\partial - \log P(O = o|C = c)}{\partial \boldsymbol{v}_w} = \frac{\partial - \log \frac{\exp\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)}{\sum_{w \in \text{Vocab}} \exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right)}}{\partial \boldsymbol{v}_w}$$

$$= -\left(\frac{\partial \log \exp\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)}{\partial \boldsymbol{v}_w} - \frac{\partial \log \sum_{w \in \text{Vocab}} \exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right)}{\partial \boldsymbol{v}_w}\right)$$

If $w = o$:

$$= -\boldsymbol{v}_c + \frac{\exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right)}{\sum_{w \in \text{Vocab}} \exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right)} \boldsymbol{v}_c$$

$$= (P(O = o|C = c) - 1) \boldsymbol{v}_c$$

If $w \neq o$:

$$= 0 + \frac{\exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right)}{\sum_{w \in \text{Vocab}} \exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right)} \boldsymbol{v}_c$$

$$= P(O = o|C = c) \boldsymbol{v}_c$$

Thus, in summary:

$$\frac{\partial \boldsymbol{J}_{\text{ naive-softmax }}(v_c, o, U)}{\partial \boldsymbol{v}_w} = (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \boldsymbol{v}_c$$

(d) The sigmoid function is given by Equation 3, Please compute the derivative of $\sigma(x)$ with respect to $\boldsymbol{x}$, where $\boldsymbol{x}$ is a vector

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

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

$$= \frac{e^x}{(e^x + 1)^2}$$

$$= \sigma(x)(1 - \sigma(x))$$

(e) 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 $w_1, w_2, \ldots, w_K$ and their outside vectors as $\boldsymbol{u}_1, ..., \boldsymbol{u}_K$. Note that $o \notin w_1, ..., w_K$. For a center word c and an outside word o, the negative sampling loss function is given by:

$$\boldsymbol{J}_{\text{ neg-sample }}(v_c, o, U) = -\log\left(\sigma\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)\right) - \sum_{k=1}^{K} \log\left(\sigma\left(-\boldsymbol{u}_k^T \boldsymbol{v}_c\right)\right) \tag{5}$$

for a sample $w_1, ...w_K$, where $\sigma(\cdot)$ is the sigmoid function. [3]

Please repeat parts (b) and (c), computing the partial derivatives of $J_{\textbf{ neg-sample }}$ with respect to $\boldsymbol{v}_c$, with respect to $\boldsymbol{u}_o$, and with respect to a negative sample $\boldsymbol{u}_k$. Please write your answers in terms of the vectors $\boldsymbol{v}_c$, $\boldsymbol{u}_o$, and $\boldsymbol{u}_k$, where $k \in [1, K]$. After you've done this, describe with one sentence why this loss function is much more efficient to compute than the naive-softmax loss. Note, you should be able to use your solution to part (d) to help compute the necessary gradients here.

1. $\boldsymbol{J}_{\text{ neg-sample }}(v_c, o, U)$ with respect to $\boldsymbol{v}_c$

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{ neg-sample }}(v_c, o, U)}{\partial \boldsymbol{v}_c} &= \frac{\partial - \log\left(\sigma\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)\right) - \sum_{k=1}^{K} \log\left(\sigma\left(-\boldsymbol{u}_k^T \boldsymbol{v}_c\right)\right)}{\partial \boldsymbol{v}_c} \\
&= -\frac{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)(1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\sigma \boldsymbol{u}_o^T \boldsymbol{v}_c} \frac{\partial \boldsymbol{u}_o^T \boldsymbol{v}_c}{\partial \boldsymbol{v}_c} - \sum_{k=1}^{K} \frac{\partial \log\left(\sigma\left(-\boldsymbol{u}_k^T \boldsymbol{v}_c\right)\right)}{\partial \boldsymbol{v}_c} \\
&= -\left(1 - \sigma\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)\right) \boldsymbol{u}_o + \sum_{k=1}^{K} \left(1 - \sigma\left(-\boldsymbol{u}_k^T \boldsymbol{v}_c\right)\right) \boldsymbol{u}_k
\end{aligned}$$

[3]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.

2. $\boldsymbol{J}_{\text{ neg-sample }}(v_c, o, U)$ with respect to $\boldsymbol{u}_o$

$$
\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{ neg-sample }}(v_c, o, U)}{\partial \boldsymbol{u}_o} = \frac{\partial - \log\left(\sigma\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)\right) - \sum_{k=1}^{K} \log\left(\sigma\left(-\boldsymbol{u}_k^T \boldsymbol{v}_c\right)\right)}{\partial \boldsymbol{u}_o} \\
= \frac{\partial\left(-\log\left(\sigma\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)\right)\right)}{\partial \boldsymbol{u}_o} \\
= -\left(1 - \sigma\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)\right) \boldsymbol{v}_c
\end{aligned}
$$

3. $\boldsymbol{J}_{\text{ neg-sample }}(v_c, o, U)$ with respect to $\boldsymbol{u}_k$

$$
\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{ neg-sample }}(v_c, o, U)}{\partial \boldsymbol{u}_k} = \frac{\partial - \log\left(\sigma\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)\right) - \sum_{k=1}^{K} \log\left(\sigma\left(-\boldsymbol{u}_k^T \boldsymbol{v}_c\right)\right)}{\partial \boldsymbol{u}_k} \\
= \frac{\partial\left(-\log\left(\sigma\left(-\boldsymbol{u}_k^T \boldsymbol{v}_c\right)\right)\right)}{\partial \boldsymbol{u}_k} \\
= \left(1 - \sigma\left(-\boldsymbol{u}_k^T \boldsymbol{v}_c\right)\right) \boldsymbol{v}_c
\end{aligned}
$$

(f) Suppose the center word is $c = w_t$ and the context window is $[w_{tm}, \ldots, w_{t1}, w_t, w_{t+1}, \ldots, w_{t+m}]$, where m is the context window size. Recall that for the skip-gram version of word2vec, the total loss for the context window is:

$$
\boldsymbol{J}_{\text{skip-gram}}\left(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U}\right) = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \boldsymbol{J}\left(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U}\right) \tag{6}
$$

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

Write down three partial derivatives:

1. $\partial \boldsymbol{J}_{\text{ skip-gram }}\left(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U}\right) / \partial \boldsymbol{U}$
2. $\partial \boldsymbol{J}_{\text{ skip-gram }}\left(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U}\right) / \partial \boldsymbol{v}_c$
3. $\partial \boldsymbol{J}_{\text{skip-gram}}\left(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U}\right) / \partial \boldsymbol{v}_w$ when $w \neq c$

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

1.

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

2.

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

3. when $w \neq c$

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