CS 224N - Assignment 2

Ilya Kachan

July 8, 2021

1 Written: Understanding word2vec

(a) $\mathbf{y}$ is the vector of true probabilities y_k of a word k being in the context of the fixed word c, therefore $y_k = 0$ if k is not o and $y_o = 1$. In view of these we obtain

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

(b) First, let's obtain an explicit expression for cross-entropy loss:

$$\mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U}) = -\log P(O = o | C = c) =$$

$$= -\mathbf{u}_o^T \mathbf{v}_c + \log \left(\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c) \right).$$

And now compute the derivative:

$$\frac{\partial \mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{v}_c} = -\mathbf{u}_o + \left(\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c) \right)^{-1} \frac{\partial}{\partial \mathbf{v}_c} \sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c) =$$

$$= -\mathbf{u}_o + \left(\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c) \right)^{-1} \sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c) \mathbf{u}_w =$$

$$= -\mathbf{u}_o + \sum_{w \in Vocab} \frac{\exp(\mathbf{u}_w^T \mathbf{v}_c)}{\sum\limits_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \mathbf{u}_w = -\mathbf{u}_o + \sum_{w \in Vocab} softmax(\mathbf{U}^T \mathbf{v}_c)_w \mathbf{u}_w =$$

$$= -\mathbf{u}_o + \mathbf{U} \cdot softmax(\mathbf{U}^T \mathbf{v}_c) = \mathbf{U}(\hat{\mathbf{y}} - \mathbf{y}),$$

where $\mathbf{y}$ is the vector with all zero components except of o-th one that is equal to 1 (true distribution) and $\hat{\mathbf{y}} = \text{softmax}(\text{U}^T \mathbf{v}_c)$ is the vector with components $\frac{\exp(\mathbf{u}_w^T \mathbf{v}_c)}{\sum\limits_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)}$, $w \in Vocab$ (i.e. the predicted distribution).

(c) Consider first the case when $w \neq o$

$$\frac{\partial \mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_w} = \left(\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c) \right)^{-1} \frac{\partial}{\partial \mathbf{u}_w} \sum_{w' \in Vocab} \exp(\mathbf{u}_{w'}^T \mathbf{v}_c) =$$

$$= \frac{\exp(\mathbf{u}_w^T \mathbf{v}_c)}{\sum\limits_{w' \in Vocab} \exp(\mathbf{u}_{w'}^T \mathbf{v}_c)} \mathbf{v}_c = \hat{y}_w \cdot \mathbf{v}_c = (\hat{y}_w - y_w) \cdot \mathbf{v}_c$$

Analogiously in case $w = o$ we obtain

$$\frac{\partial \mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_o} = -\mathbf{v}_c + \left(\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c) \right)^{-1} \frac{\partial}{\partial \mathbf{u}_o} \sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c) =$$

$$= -\mathbf{v}_c + \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum\limits_{w \in Vocab} \exp(\mathbf{u}_o^T \mathbf{v}_c)} \mathbf{v}_c = (\hat{y}_o - 1) \cdot \mathbf{v}_c = (\hat{y}_o - y_o) \cdot \mathbf{v}_c$$

(d) It can be easily seen that

$$\frac{\partial \mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{U}} = \left[\frac{\partial \mathbf{J}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_1}, \frac{\partial \mathbf{J}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_2}, \dots, \frac{\partial \mathbf{J}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_{|Vocab|}} \right],$$

where $\frac{\partial \mathbf{J}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_w} = -\hat{\mathbf{y}}_w \cdot \mathbf{v}_c$ if $w \neq o$ and $\frac{\partial \mathbf{J}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_o} = -(1 + \hat{\mathbf{y}}_o) \cdot \mathbf{v}_c$.

(e) Let's compute the derivative of the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$:

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

$$= \frac{1}{e^x + 1} \cdot \frac{1}{1+e^{-x}} = \sigma(-x)\sigma(x).$$

(f) Consider negative sampling loss function

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

It's computationally much cheeper than the naive softmax loss since the summation inside its formula is done over the small set of K negative samples instead of the whole vocabulary $Vocab$.

Let's compute its derivatives in the assumption that all $\mathbf{u}_k$ are different:

$$\frac{\partial \mathbf{J}_{neg-sample}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{v}_c} =$$

$$= -\sigma^{-1}(\mathbf{u}_o^T \mathbf{v}_c)\sigma'(\mathbf{u}_o^T \mathbf{v}_c)\mathbf{u}_o - \sum_{k=1}^{K} \sigma^{-1}(-\mathbf{u}_k^T \mathbf{v}_c)\sigma'(-\mathbf{u}_k^T \mathbf{v}_c)(-\mathbf{u}_k) =$$

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

$$\frac{\partial \mathbf{J}_{neg-sample}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_o} = -\sigma^{-1}(\mathbf{u}_o^T \mathbf{v}_c)\sigma'(\mathbf{u}_o^T \mathbf{v}_c)\mathbf{v}_c = -\sigma(-\mathbf{u}_o^T \mathbf{v}_c)\mathbf{v}_c.$$

$$\frac{\partial \mathbf{J}_{neg-sample}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_k} = -\sigma^{-1}(-\mathbf{u}_k^T \mathbf{v}_c)\sigma'(-\mathbf{u}_k^T \mathbf{v}_c)(-\mathbf{v}_c) = \sigma(\mathbf{u}_k^T \mathbf{v}_c)\mathbf{v}_c.$$

(g) Now let's compute the last derivative without the assumption that all $\mathbf{u}_k$ are different:

$$\mathbf{J}_{neg-sample}(\mathbf{v}_c, o, \mathbf{U}) =$$

$$= -\log(\sigma(\mathbf{u}_o^T \mathbf{v}_c)) - \sum_{k' :\, u_{k'} = u_k} \log(\sigma(-\mathbf{u}_{k'}^T \mathbf{v}_c)) - \sum_{k' :\, u_{k'} \neq u_k} \log(\sigma(-\mathbf{u}_{k'}^T \mathbf{v}_c)) =$$

$$= -\log(\sigma(\mathbf{u}_o^T \mathbf{v}_c)) - n_k \log(\sigma(-\mathbf{u}_k^T \mathbf{v}_c)) - \sum_{k' :\, u_{k'} \neq u_k} \log(\sigma(-\mathbf{u}_{k'}^T \mathbf{v}_c)),$$

where n_k is the number of words among $w_1, w_2, \ldots, w_K$ equal to w_k. We get

$$\frac{\partial \mathbf{J}_{neg-sample}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_k} = -n_k\sigma^{-1}(-\mathbf{u}_k^T \mathbf{v}_c)\sigma'(-\mathbf{u}_k^T \mathbf{v}_c)(-\mathbf{v}_c) = n_k\sigma(\mathbf{u}_k^T \mathbf{v}_c)\mathbf{v}_c.$$

(h) Compute the derivatives of skip-gram loss:

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

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

$$\frac{\partial \mathbf{J}_{skip-gram}(\mathbf{v}_c, w_{t-m}, \ldots, w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_w} = 0, \quad w \neq c.$$