CS 224n
Problem Set #2 Solutions: word2vec

LA DUC CHINH

Due Wednesday, Nov 14 at 11:59 pm on Gradescope.

Written: Understanding word2vec

(a) We have

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

(b) We have

$$\begin{aligned}
\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial v_c} &= -u_o + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} \sum_{w \in Vocab} exp(u_w^T v_c) u_w & (2) \\
&= -u_o + \sum_{w \in Vocab} \frac{exp(u_w^T v_c)}{\sum_{w \in Vocab} exp(u_w^T v_c)} u_w & (3) \\
&= -u_o + \sum_{w \in Vocab} P(O = w | C = c) u_w & (4) \\
&= -u_o + \sum_{w \in Vocab} \hat{y}_w u_w & (5)
\end{aligned}$$

Let $\hat{y}_w = y_w + (\hat{y}_w - y_w)$ and we have $\sum_{w \in Vocab} y_w u_w = \sum_{w \in Vocab} 1\{w = o\} u_w u_o$

$$
\begin{aligned}
\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial v_c} &= -u_o + \sum_{w \in Vocab} \hat{y_w} u_w && (6)\\
&= -u_o + \sum_{w \in Vocab} [y_w + (\hat{y}_w - y_w)] u_w && (7)\\
&= -u_o + \sum_{w \in Vocab} y_w u_w + \sum_{w \in Vocab} (\hat{y}_w - y_w) u_w && (8)\\
&= -u_o + u_o + \sum_{w \in Vocab} (\hat{y}_w - y_w) u_w && (9)\\
&= \sum_{w \in Vocab} (\hat{y}_w - y_w) u_w && (10)\\
&= \sum_{j=1}^{|V|} (\hat{y_w} - y_w) u_w \qquad = (\hat{y_1} - y_1)u_1 + (\hat{y_2} - y_2)u_2 + ... + (\hat{y} && (11)\\
&= \begin{pmatrix} \hat{y_1} - y_1 & \hat{y_2} - y_2 & \cdots & \hat{y_{|V|}} - y_{|V|} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{|V|} \end{pmatrix} && (12)\\
&= (\hat{y} - y) U && (13)
\end{aligned}
$$

(c) partial derivatives outside word when $w = o$
Similar to (b) we have

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_o} = -(1 - \hat{y}) \boldsymbol{v}_c = (\hat{y}_{w=o} - y_{y=o}) \boldsymbol{v}_c \tag{14}$$

When $w \neq o$

$$
\begin{aligned}
\frac{\partial J}{\partial \boldsymbol{u}_w} &= -\frac{\partial \log \hat{y}}{\partial \boldsymbol{u}_w} && (15)\\
&= -\frac{\partial \log \hat{y}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \boldsymbol{u}_w^T \boldsymbol{v}_c} \frac{\partial \boldsymbol{u}_w^T \boldsymbol{v}_c}{\partial \boldsymbol{u}_w} && (16)\\
&= -\frac{1}{\hat{y}} \frac{-\exp(\boldsymbol{u}_o^T \boldsymbol{v}_c) \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\left(\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)\right)^2} \boldsymbol{v}_c && (17)\\
&= \frac{1}{\hat{y}} \hat{y} \hat{y}_{w, w \neq o} \boldsymbol{v}_c && (18)\\
&= \hat{y}_{w, w \neq o} \boldsymbol{v}_c && (19)\\
&= (\hat{y}_{w, w \neq o} - y_{w, w \neq o}) \boldsymbol{v}_c && (20)
\end{aligned}
$$

Therefore we have

$$\frac{\partial J}{\partial \boldsymbol{u}_w} = (\hat{y}_w - y_w) \boldsymbol{v}_c \tag{21}$$

(d) Sigmoid

$$\frac{d\sigma}{d\boldsymbol{x}} = \sigma(1-\sigma) \tag{22}$$

(e) Negative Sampling loss

$$\boldsymbol{J}_{neg-sample}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log(\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)) - \sum_{k=1}^{K}\log(\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c)) \tag{23}$$

Derivative $\boldsymbol{v}_c$

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{v}_c} = -\frac{1}{\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)}\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)(1-\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c))\boldsymbol{u}_o - \sum_{k=1}^{K}\frac{1}{\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c)}\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c)(1-\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c))(-\boldsymbol{u}_k) \tag{24}$$

$$= -(1-\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c))\boldsymbol{u}_o + \sum_{k=1}^{K}(1-\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c))\boldsymbol{u}_k \tag{25}$$

Derivative $\boldsymbol{u}_o$

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_o} = -\frac{1}{\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)}\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)(1-\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c))\boldsymbol{v}_c \tag{26}$$

$$= -(1-\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c))\boldsymbol{v}_c \tag{27}$$

Derivative $\boldsymbol{u}_k$

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_k} = \sum_{k=1}^{K} -\frac{1}{\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c)}\sigma(\boldsymbol{u}_k^T\boldsymbol{v}_c)(1-\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c))(-\boldsymbol{v}_c) \tag{28}$$

$$= \sum_{k=1}^{K}(1-\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c))\boldsymbol{v}_c \tag{29}$$

This loss function is more efficient because sigmoid function has less computation cost than softmax function.

(f) Skip-gram

$$\frac{\partial J_{\text{skip-gram}}}{\partial U} = \sum_{\substack{-m\le j\le m \\ j\ne 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U} \tag{30}$$

$$\frac{\partial J_{\text{skip-gram}}}{\partial v_c} = \sum_{\substack{-m\le j\le m \\ j\ne 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c} \tag{31}$$

$$\frac{\partial J_{\text{skip-gram}}}{\partial v_w} = 0 \tag{32}$$