Shanghai Jiao Tong University

Natural Language Processing with Deep Learning: HW 2

Author: *LIUQiaoan 520030910220*

Course: *CS224n: Natural Language Processing with Deep Learning, Stanford, Winter 2021*
Date: *October 28, 2022*

(a)

Since o is the only true outside word, then

$$\forall w \in \text{Vocab}, y_w = \begin{cases} 1, & w = o \\ 0, & \text{o.w.} \end{cases}$$

So for all $o \in \text{Vocab}$, we have

$$\begin{aligned} -\sum_{w\in\text{Vocab}} y_w \log(\hat{y}_w) &= -\sum_{w\in\text{Vocab}} 1\{y_w = 1\} \log(\hat{y}_w) \\ &= -\log(\hat{y}_o) \\ &= -\log P(O = o | C = c) \\ &= \boldsymbol{J}_{\text{naive\_softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) \end{aligned}$$

(b)

$$\begin{aligned} \frac{\partial \boldsymbol{J}_{\text{naive\_softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} &= \frac{\partial \left\{-\boldsymbol{u}_o^T \boldsymbol{v}_c + \log \sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)\right\}}{\partial \boldsymbol{v}_c} \\ &= -\boldsymbol{u}_o + \frac{\sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)\boldsymbol{u}_w}{\sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \\ &= -\boldsymbol{u}_o + \sum_{w\in\text{Vocab}} P(O = w|C = c)\boldsymbol{u}_w \\ &= -\boldsymbol{u}_o + \sum_{w\in\text{Vocab}} \hat{y}_w \boldsymbol{u}_w \\ &= \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}) \qquad \text{(This is such a clean form!)} \end{aligned}$$

(c)

When $w = o$,

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

When $w \neq o$,

$$
\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{naive\_softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_w} &= \frac{\partial \left\{-\boldsymbol{u}_o^T \boldsymbol{v}_c + \log \sum_{x \in \text{Vocab}} \exp\left(\boldsymbol{u}_x^T \boldsymbol{v}_c\right)\right\}}{\partial \boldsymbol{u}_x} \\
&= \frac{\exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right) \boldsymbol{v}_c}{\sum_{x \in \text{Vocab}} \exp\left(\boldsymbol{u}_x^T \boldsymbol{v}_c\right)} \\
&= \hat{y}_w \boldsymbol{v}_c
\end{aligned}
$$

(d)

$$
\frac{\partial \boldsymbol{J}_{\text{naive\_softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{U}} = \left[\frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_1}, \quad \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_2}, \quad \ldots, \quad \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_{|Vocab|}}\right]
$$

(e)

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

(f)

$$
\frac{\partial \boldsymbol{J}_{\text{neg\_sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = -(1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) \boldsymbol{u}_o + \sum_{k=1}^{K} (1 - \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \boldsymbol{u}_k
$$

$$
\frac{\partial \boldsymbol{J}_{\text{neg\_sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = -(1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) \boldsymbol{v}_c
$$

$$
\frac{\partial \boldsymbol{J}_{\text{neg\_sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_k} = (1 - \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \boldsymbol{v}_c
$$

Compare (b) and (c) with (f), we can see that (b) requires matrix multiplication, while (f) only needs inner product.

(g)

$$\frac{\partial \boldsymbol{J}_{\text{neg\_sample}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_k} = \sum_{w=1}^{K} 1\{\boldsymbol{u}_w = \boldsymbol{u}_k\}(1 - \sigma(-\boldsymbol{u}_w^T \boldsymbol{v}_c))\boldsymbol{v}_c$$

(h)

i.

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

ii.

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

iii.

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