Assignment 2: word2vec

1. Written: Understanding word2vec

(a) True output vector y is 1 in o^{th} position and 0 else where. So the cross-entropy $-\sum_{w\in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$

(b) $J_{navie-softmax}(v_c, o, U) = -u_o^T v_c + \log\sum_{w\in Vocab}\exp(u_w^T v_c)$

$$\begin{aligned}
\frac{\partial J}{\partial v_c} &= -u_o + \frac{\frac{\partial}{\partial v_c}\sum_{w\in Vocab}\exp(u_w^T v_c)}{\sum_{w\in Vocab}\exp(u_w^T v_c)} \\
&= -u_o + \frac{\sum_{w\in Vocab}\exp(u_w^T v_c)u_w}{\sum_{w\in Vocab}\exp(u_w^T v_c)} \\
&= -u_o + \sum_{\hat{w}\in Vocab}\frac{\exp(u_{\hat{w}}^T v_c)}{\sum_{w\in Vocab}\exp(u_w^T v_c)}u_{\hat{w}} \\
&= -Uy + \sum_{\hat{w}\in Vocab} u_{\hat{w}}\hat{y}_{\hat{w}} \\
&= -Uy + U\hat{y} \\
&= U(\hat{y} - y)
\end{aligned}$$

$$where\ U \in \mathbb{R}^{d\times|V|},\ y \in \mathbb{R}^{|V|\times 1},\ \hat{y} \in \mathbb{R}^{|V|\times 1}$$

(c) $J_{navie-softmax}(v_c, o, U) = -u_o^T v_c + \log\sum_{w\in Vocab}\exp(u_w^T v_c)$

$$\begin{aligned}
&(w \neq o) \\
&\quad \frac{\partial J}{\partial u_w} = \frac{\frac{\partial}{\partial u_w}\sum_{w\in Vocab}\exp(u_w^T v_c)}{\sum_{w\in Vocab}\exp(u_w^T v_c)} \\
&\quad\quad\quad = \frac{\exp(u_w^T v_c)v_c}{\sum_{w\in Vocab}\exp(u_w^T v_c)} \\
&\quad\quad\quad = \hat{y}_w v_c \\
&(w = o) \\
&\quad \frac{\partial J}{\partial u_w} = -v_c + \frac{\frac{\partial}{\partial u_w}\sum_{w\in Vocab}\exp(u_w^T v_c)}{\sum_{w\in Vocab}\exp(u_w^T v_c)} \\
&\quad\quad\quad = -v_c + \frac{\exp(u_w^T v_c)v_c}{\sum_{w\in Vocab}\exp(u_w^T v_c)} \\
&\quad\quad\quad = (\hat{y}_w - 1)v_c
\end{aligned}$$

$$Then\ \frac{\partial J(v_c, o, U)}{\partial U} = v_c(\hat{y} - y)^T$$

(d)

$$
\begin{aligned}
\frac{\partial \sigma(x_i)}{\partial x_i} &= \sigma(x_i) - \frac{e^{x_i}\frac{\partial}{\partial x_i}(e^{x_i}+1)}{(e^{x_i}+1)^2} \\
&= \sigma(x_i) - \frac{e^{2x_i}}{(e^{x_i}+1)^2} \\
&= \sigma(x_i) - \sigma^2(x_i)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\sigma(x)}{\partial x} &= [\frac{\partial \sigma(x_j)}{\partial x_i}]_{d\times d} \\
&= \begin{bmatrix} \sigma'(x_1) & 0 & \cdots & 0 \\ 0 & \sigma'(x_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \sigma'(x_d) \end{bmatrix} \\
&= diag(\sigma'(x))
\end{aligned}
$$

(e)

$$
J_{neg-sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))
$$

$$
\begin{aligned}
&Respect\ to\ v_c: \\
&\frac{\partial J}{\partial v_c} = -\frac{\frac{\partial}{\partial v_c}\sigma(u_o^T v_c)}{\sigma(u_o^T v_c)} - \sum_{k=1}^{K} \frac{\frac{\partial}{\partial v_c}\sigma(-u_k^T v_c)}{\sigma(-u_k^T v_c)} \\
&\quad = -\frac{\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))u_o}{\sigma(u_o^T v_c)} + \sum_{k=1}^{K} \frac{\sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c))u_k}{\sigma(-u_k^T v_c)} \\
&\quad = (\sigma(u_o^T v_c) - 1)u_o + \sum_{k=1}^{K} \sigma(u_k^T v_c)u_k
\end{aligned}
$$

$$
\begin{aligned}
&Respect\ to\ u_o: \\
&\frac{\partial J}{\partial u_o} = -\frac{\frac{\partial}{\partial u_o}\sigma(u_o^T v_c)}{\sigma(u_o^T v_c)} \\
&\quad = (\sigma(u_o^T v_c) - 1)v_c
\end{aligned}
$$

$$
\begin{aligned}
&Respect\ to\ u_k: \\
&\frac{\partial J}{\partial u_k} = -\frac{\partial}{\partial u_k}\sum_{k=1}^{K}\log(\sigma(-u_k^T v_c)) \\
&\quad = (1-\sigma(-u_k^T v_c))v_c \\
&\quad = \sigma(u_k^T v_c)v_c,\ for\ k = 1, 2, \ldots, K
\end{aligned}
$$

(f)

$$\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial U} = \sum_{-m \le j \le m,\ j \neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

$$\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial v_c} = \sum_{-m \le j \le m,\ j \neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

$$\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial v_w} = 0$$

The plot of my training