Derivation of Word2Vec Models

Lifu Huang

Feb. 2016

1 CBOW Model

Given a word windows of length $C+1$, assume that the index of the the center word in the window is w_O, the context words are $w_1, w_2, .., w_C$.

1.1 Loss Function

$$\begin{aligned} E &= -\log p(w_O|w_1, w_2, .., w_C) \\ &= -\log \left(\frac{e^{u_{w_O}}}{\sum_{k=1}^{V} e^{u_k}} \right) \\ &= -u_{w_O} + \log \sum_{k=1}^{V} e^{u_k} \end{aligned} \tag{1}$$

1.2 Updating W'

$$\begin{aligned} \delta'_i &= \frac{\partial E}{\partial u_i} \\ &= -t_i + \frac{e^{u_i}}{\sum_{k=1}^{V} e^{u_k}} \\ &= y_i - t_i \end{aligned} \tag{2}$$

$$\delta' = \nabla_u E = y - t \tag{3}$$

$$\begin{aligned} \frac{\partial E}{\partial W'_{ij}} &= \sum_{k=1}^{V} \left(\frac{\partial E}{\partial u_k} \cdot \frac{\partial u_k}{\partial W'_{ij}} \right) \\ &= \sum_{k=1}^{V} \left(\delta'_k \cdot \frac{\partial u_k}{\partial W'_{ij}} \right) \\ &= \delta'_k \cdot h_j \end{aligned} \tag{4}$$

$$\nabla_{W'} E = \delta' \cdot h^T \tag{5}$$

Updating Rules

$$v'_{i,new} = v'_{i,old} - \eta \cdot \delta'_i \cdot h \quad \text{for i = 1, 2, .., V} \tag{6}$$

1.3 Updating W

$$\begin{aligned} \delta_i &= \frac{\partial E}{\partial h_i} \\ &= \sum_{j=1}^{V} \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial h_i} \\ &= \sum_{j=1}^{V} \delta'_j \cdot W'_{ji} \end{aligned} \tag{7}$$

$$\delta = (W')^T \cdot \delta' \tag{8}$$

$$\begin{aligned} \frac{\partial E}{W_{ij}} &= \sum_{k=1}^{N} \frac{\partial E}{h_k} \cdot \frac{\partial h_k}{\partial W_{ij}} \\ &= \sum_{k=1}^{N} \delta_k \cdot \frac{\partial}{\partial W_{ij}} \left(\frac{1}{C} \sum_{c=1}^{C} W_{k\cdot} \cdot x_{w_c} \right) \\ &= \frac{1}{C} \cdot \delta_i \cdot \sum_{c=1}^{C} x_{w_c,j} \end{aligned} \tag{9}$$

$$\nabla_W E = \frac{1}{C} \cdot \delta \cdot \left(\sum_{c=1}^{C} x_{w_c} \right)^T \tag{10}$$

Updating Rules

$$v_{w_i,new} = v_{w_i,old} - \eta \cdot \frac{1}{C} \cdot \delta \quad \text{for i = 1, 2, .., C} \tag{11}$$

2 Skip-Gram Model

Given a word windows of length $C+1$, assume that the index of the the center word in the window is w_I, the context words are $w_1, w_2, .., w_C$.

2.1 Loss Function

$$\begin{aligned} E &= -\log p(w_1, w_2, .., w_C | w_I) \\ &= -\log \prod_{c=1}^{C} p(w_c | w_I) \\ &= -\sum_{c=1}^{C} \log p(w_c | w_I) \\ &= -\sum_{c=1}^{C} \log \frac{e^{u_{c,w_c}}}{\sum_{k=1}^{V} e^{u_{ck}}} \end{aligned} \tag{12}$$

2.2 Updating W'

$$\begin{aligned} \delta'_{ci} &= \frac{\partial E}{\partial u_{ci}} \\ &= \frac{\partial}{\partial u_{ci}} \left(-\sum_{c=1}^{C} \log \frac{e^{u_{c,w_c}}}{\sum_{k=1}^{V} e^{u_{ck}}} \right) \\ &= -t_{ci} + \frac{e^{u_{ci}}}{\sum_{k=1}^{V} e^{u_{ck}}} \\ &= y_{ci} - t_{ci} \end{aligned} \tag{13}$$

$$\delta'_c = y_c - t_c \tag{14}$$

$$\begin{aligned} \frac{\partial E}{\partial W'_{ij}} &= \sum_{c=1}^{C} \sum_{k=1}^{V} \frac{\partial E}{\partial u_{ck}} \cdot \frac{\partial u_{ck}}{\partial W'_{ij}} \\ &= \sum_{c=1}^{C} \sum_{k=1}^{V} \delta'_{ck} \cdot \frac{\partial}{\partial W'_{ij}} \sum_{l=1}^{N} W'_{kl} \cdot h_l \\ &= \left(\sum_{c=1}^{C} \delta'_{ci} \right) \cdot h_j \\ &= \gamma_i \cdot h_j \end{aligned} \tag{15}$$

$$\nabla_{W'} E = \gamma \cdot h^T \tag{16}$$

Updating Rules

$$\boxed{v'_{i,new} = v'_{i,old} - \eta \cdot \gamma_i \cdot h \quad \text{for i = 1, 2, .., V} \tag{17}}$$

2.3 Updating W

$$\begin{aligned}
\delta_i &= \frac{\partial E}{\partial h_i} \\
&= \sum_{c=1}^{C} \sum_{j=1}^{V} \left(\frac{\partial E}{u_{cj}} \cdot \frac{\partial u_{cj}}{\partial h_i} \right) \\
&= \sum_{c=1}^{C} \sum_{j=1}^{V} \left(\delta'_{cj} \cdot W'_{j,i} \right) \\
&= \sum_{j=1}^{V} W'_{ji} \cdot \left(\sum_{c=1}^{C} \delta'_{cj} \right) \\
&= \sum_{j=1}^{V} W'_{ji} \cdot \gamma_j
\end{aligned} \tag{18}$$

$$\delta = (W')^T \cdot \gamma \tag{19}$$

$$\begin{aligned}
\frac{\partial E}{\partial W_{ij}} &= \sum_{k=1}^{N} \left(\frac{\partial E}{\partial h_k} \cdot \frac{\partial h_k}{\partial W_{ij}} \right) \\
&= \sum_{k=1}^{N} \left(\delta_k \cdot \frac{\partial h_k}{W_{ij}} \right) \\
&= \delta_i \cdot x_{w_I,j}
\end{aligned} \tag{20}$$

$$\nabla_W E = \delta \cdot x_{w_I}^T \tag{21}$$

Updating Rules

$$\boxed{v_{w_I,new} = v_{w_I,old} - \eta \cdot \delta \tag{22}}$$

3 Negative Sampling

The derivation for negative sampling is almost the same as before, except that we changed our objective function to its approximated version for better performance. As a result, all we need is to recalculate δ', which can then be plugged back into formula in CBOW or Skip-Gram model we derived before.

3.1 CBOW

$$E = -\log\sigma(u_{w_O}) - \sum_{k\sim P_n(w)} \log\sigma(-u_k) \tag{23}$$

$$\begin{aligned}\delta'_i &= \frac{\partial E}{\partial u_i} \\ &= \begin{cases}\sigma(u_i) - t_i & \text{for } i \in \{w_O\} \cup \{k \sim P_n(w)\} \\ 0 & \text{Otherwise}\end{cases}\end{aligned} \tag{24}$$

Complete Updating Rules

Let $S = \{w_O\} \cup \{k \sim P_n(w)\}$,

$$\delta'_i = \frac{\partial E}{\partial u_i} = \sigma(u_i) - t_i \quad \text{for } i \in S \tag{25}$$

$$v'_{i,new} = v'_{i,old} - \eta \cdot \delta'_i \cdot h \quad \text{for } i \in S \tag{26}$$

$$\delta = \sum_{i\in S} \delta'_i \cdot v'_i \tag{27}$$

$$v_{w_i,new} = v_{w_i,old} - \eta \cdot \frac{1}{C} \cdot \delta \quad \text{for i = 1, 2, .., C} \tag{28}$$

3.2 Skip-Gram

$$E = \sum_{c=1}^{C}\left(-\log\sigma(u_{c,w_c}) - \sum_{k\sim P_n(w)} \log\sigma(-u_{c,k})\right) \tag{29}$$

$$\begin{aligned}\delta'_{ci} &= \frac{\partial E}{\partial u_{ci}} \\ &= \begin{cases}\sigma(u_{ci}) - t_{ci} & \text{for } i \in \{w_c\} \cup \{k \sim P_n(w)\} \\ 0 & \text{Otherwise}\end{cases}\end{aligned} \tag{30}$$

Complete Updating Rules

For each context word w_c,

$$\text{let } S_c = \{w_c\} \cup \{k \sim P_n(w)\},$$

$$\delta'_{ci} = \frac{\partial E}{\partial u_{ci}} = \sigma(u_{ci}) - t_{ci} \quad \text{for } i \in S_c \tag{31}$$

$$v'_{i,new} = v'_{i,old} - \eta \cdot \delta'_{ci} \cdot h \quad \text{for } i \in S_c \tag{32}$$

$$\delta_c = \sum_{i \in S_c} \delta'_{ci} \cdot v'_i \tag{33}$$

$$v_{w_I,new} = v_{w_I,old} - \eta \cdot \delta_c \tag{34}$$