Calculations of Gradients for the Skip-gram Model

Xiuwen Liu
Department of Computer Science, Florida State University

1 The Skip-gram Model

We assume a dictionary with W words is given. For each word w_t, we like to learn an input vector representation v_{w_t} and output vector representation $v^{'}_{w_t}$ from a dataset of size T, consisting of w_1, w_2, ..., w_T by maximizing

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t), \tag{1}$$

where

$$p(w_O|w_I)=\frac{\exp(v^{'}_{w_O}\cdot v_{w_I})}{\sum_{w=1}^{W}\exp(v^{'}_{w}\cdot v_{w_I})}. \tag{2}$$

Note that w_t is the index of the word in the dictionary and it is not the word itself. In other words, the model does not depend what the actual words and it only depends on the ordering of occurrences of the words.

2 Gradient-descent Learning Rules for the Skip-gram Model

We assume the input and output vectors for the words are initialized and we like to use gradient descent optimization algorithm to optimize the objective function. We define

$$\begin{aligned}
E &= \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)\\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\left(-v^{'}_{w_{t+j}}\cdot v_{w_t}+\log\left(\sum_{w=1}^{W}\exp(v^{'}_{w}\cdot v_{w_t})\right)\right)\\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\left(\sum_{-c\leq j\leq c, j\neq 0}\left(-v^{'}_{w_{t+j}}\cdot v_{w_t}\right)\right)+2c\log\left(\sum_{w=1}^{W}\exp(v^{'}_{w}\cdot v_{w_t})\right)\right).
\end{aligned}$$

For an input vector v_{w_s}, the gradient is given as:

$$\begin{aligned}
\frac{\partial E}{\partial v_{w_{s,i}}} &= \frac{1}{T}\sum_{t=1,w_t=w_s}^{T}\sum_{-c\le j\le c, j\neq 0}\left(-v'_{w_{t+j,i}} + \frac{\sum_{w=1}^{W}\exp(v'_w\cdot v_{w_t})v'_{w_i}}{\sum_{w=1}^{W}\exp(v'_w\cdot v_{w_t})}\right)\\
&= \frac{1}{T}\sum_{t=1,w_t=w_s}^{T}\sum_{-c\le j\le c, j\neq 0}\left(-v'_{w_{t+j,i}} + \sum_{w=1}^{W}\frac{\exp(v'_w\cdot v_{w_t})v'_{w_i}}{\sum_{w'=1}^{W}\exp(v'_{w'}\cdot v_{w_t})}\right)\\
&= \frac{1}{T}\sum_{t=1,w_t=w_s}^{T}\sum_{-c\le j\le c, j\neq 0}\left(-v'_{w_{t+j,i}} + \sum_{w=1}^{W}p(w|w_t)v'_{w_i}\right)\\
&= \frac{1}{T}\sum_{t=1,w_t=w_s}^{T}\left(\left(\sum_{-c\le j\le c, j\neq 0}-v'_{w_{t+j,i}}\right) + 2c\times\sum_{w=1}^{W}p(w|w_t)v'_{w_i}\right).
\end{aligned}$$

Note that we need to sum over all the occurrences of w_s in the dataset. The updating rule for the input vector of word w_s is

$$v_{w_{s,i}}{}^{\text{new}} = v_{w_{s,i}}{}^{\text{old}} + \eta\times\frac{1}{T}\sum_{t=1,w_t=w_s}^{T}\sum_{-c\le j\le c, j\neq 0}\left(v'_{w_{t+j,i}} - \sum_{w=1}^{W}p(w|w_t)v'_{w_i}\right). \tag{3}$$

Therefore, for the entire vector, we have:

$$v_{w_s}{}^{\text{new}} = v_{w_s}{}^{\text{old}} + \eta\times\frac{1}{T}\sum_{t=1,w_t=w_s}^{T}\sum_{-c\le j\le c, j\neq 0}\left(v'_{w_{t+j}} - \sum_{w=1}^{W}p(w|w_t)v'_{w}\right). \tag{4}$$

Can we interpret the result intuitively?

For an output vector v'_{w_s}, the gradient is given as:

$$\begin{aligned}
\frac{\partial E}{\partial v_{w'_{s,i}}} &= \frac{1}{T}\sum_{t=1}^{T}\left(\left(\sum_{-c\le j\le c, j\neq 0, w_{t+j}=w_s}-v_{w_{t,i}}\right) + 2c\frac{\exp(v'_{w_s}\cdot v_{w_t})v_{w_{t,i}}}{\sum_{w'=1}^{W}\exp(v'_{w'}\cdot v_{w_t})}\right)\\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\left(\sum_{-c\le j\le c, j\neq 0, w_{t+j}=w_s}-v_{w_{t,i}}\right) + 2c\times p(w_s|w_t)v_{w_{t,i}}\right).
\end{aligned}$$

Note that we need to sum over all the occurrences of w_s in the contexts. The updating rule for the output vector of w_s is

$$v'_{w_{s,i}}{}^{\text{new}} = v'_{w_{s,i}}{}^{\text{old}} + \eta\times\frac{1}{T}\sum_{t=1}^{T}\left(\left(\sum_{-c\le j\le c, j\neq 0, w_{t+j}=w_s}v_{w_{t,i}}\right) - 2c\times p(w_s|w_t)v_{w_{t,i}}\right). \tag{5}$$

Therefore, for the entire vector, we have:

$$v'_{w_s}{}^{\text{new}} = v'_{w_s}{}^{\text{old}} + \eta\times\frac{1}{T}\sum_{t=1}^{T}\left(\left(\sum_{-c\le j\le c, j\neq 0, w_{t+j}=w_s}v_{w_t}\right) - 2c\times p(w_s|w_t)v_{w_t}\right). \tag{6}$$

Also, can we interpret the result intuitively?

3 Experimental Results

Given the following paragraph:

"Every person had a star, every star had a friend, and for every person carrying a star there was someone else who reflected it, and everyone carried this reflection like a secret confidante in the heart."

Here the dictionary consists of the 26 unique words in the paragraph. If we apply the skip-gram algorithm on it with $c = 2$ and vector length of 10, which word should have the most similar input vector representation to that of **person**? Which word should have the most similar output vector representation to that of **person**? Which word should have the most similar output vector representation to the input vector representation of **person**?