Note: Neural Word Embedding as Implicit Matrix Factorization

Alicia Tsai

April 2020

This note is a summary of the paper Neural Word Embedding as Implicit Matrix Factorization [1]. All typos are on me.

1 Skip-Gram with Negative Sampling (SGNS)

1.1 Notation

The skip-gram model assumes a corpus of words $w \in V_W$ and their context $c \in V_C$, where V_W and V_C are the word and context vocabularies. The words typically come from un-annotated corpora of words $w_1, w_2, \cdots, w_n$, and the context for word w_i are the words surrounding it in an L-sized window $w_{i-L}, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{i+L}$. The collection of observed word and context pairs are denoted as D. We use $\#(w,c)$ to denote the number of times the pair (w,c) appears in D. Similarly, $\#(w) = \sum_{c' \in V_c} \#(w,c')$ and $\#(c) = \sum_{w' \in V_W} \#(w',c)$ are the number of times w and c occurred in D, respectively.

Each word w and context c is associated with a vector $\vec{w} \in \mathbb{R}^d, \vec{c} \in \mathbb{R}^d$, where d is the embedding's dimension. These vectors are parameters to be learned. We can put the vectors into a matrix $W(C)$ with dimension $|V_W| \times d$ $(|V_C| \times d)$ where each row W_i (C_i) refers to the vector representation of the ith word (context) in the corresponding vocabularies.

1.2 SGNS's Objective

Consider a word-context pair (w,c), the probability that (w,c) came from the data is modeled as:

$$P(D=1|w,c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1+e^{-\vec{w} \cdot \vec{c}}} \tag{1}$$

Similarly, the probability that (w,c) did not come from the data is modeled as:

$$\begin{aligned} P(D=0|w,c) &= 1 - P(D=1|w,c) \\ &= 1 - \sigma(\vec{w} \cdot \vec{c}) = \sigma(-\vec{w} \cdot \vec{c}) \end{aligned} \tag{2}$$

The SGNS's objective tries to maximize $P(D=1|w,c)$ for observed (w,c) pairs and to maximize $P(=0|w,c)$ for randomly sampled "negative" examples, under the assumption that randomly selecting a context for a given word is likely to result in an unobserved (w,c) pair. For a single (w,c) observation, the objective function is then:

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c_N})] \tag{3}$$

where k is the number of "negative" samples and c_N is the sampled negative context, drawn according to the empirical unigram distribution $P_D(c) = \frac{\#(w,c)}{|D|}$. Finally, the global objective function sums over the observed (w,c) pairs in the corpus:

$$L = \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c) \Big(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c_N})] \Big) \tag{4}$$

Optimizing this objective makes observed word-context pairs have similar embeddings, while scattering unobserved pairs.

2 SGNS as Implicit Matrix Factorization

2.1 Characterizing Implicit Matrix

SGNS embeds both words and their contexts into a low-dimensional space $\mathbb{R}^d$, resulting in the word and context matrices W and C. Consider the product $W \cdot C^T = M$, the SGNS can be described as *factorizing* an implicit matrix M of dimensions $|V_W| \times |V_C|$ into two smaller matrices. Each entry M_{ij} in M corresponds to the dot product $W_i \cdot C_j^T = \vec{w_i} \cdot \vec{c_j}$. In other words, each entry contains a quantity $f(w,c)$ reflecting the strength of association between that particular word-context (w,c) pair.

Consider the global objective (equation 4) above. For sufficiently large dimension d that allows for a perfect reconstruction of M, each product $\vec{w} \cdot \vec{c}$ can assume a value independently of the others. Under these conditions, we can treat the objective L as a function of independent $\vec{w} \cdot \vec{c}$ terms, and find the values of these terms that maximize it.

We start from rewriting equation 4:

$$
\begin{aligned}
L &= \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c)\Big(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c_N})]\Big) \\
&= \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c)\Big(\log \sigma(\vec{w} \cdot \vec{c})\Big) + \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c)\Big(k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c_N})]\Big) \\
&= \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c)\Big(\log \sigma(\vec{w} \cdot \vec{c})\Big) + \sum_{w \in V_W} \#(w)\Big(k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c_N})]\Big)
\end{aligned}
\tag{5}
$$

The expectation can be expressed explicitly:

$$
\begin{aligned}
\mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c_N})] &= \sum_{c_N \in V_C} \frac{\#(c_N)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c_N}) \\
&= \frac{\#(c)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c}) + \sum_{c_N \in V_C \setminus \{c\}} \frac{\#(c_N)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c_N})
\end{aligned}
\tag{6}
$$

Combining equation 5 and 6, we get:

$$
\begin{aligned}
L = &\sum_{w \in V_W} \sum_{c \in V_C} \#(w,c)\Big(\log \sigma(\vec{w} \cdot \vec{c})\Big) + \sum_{w \in V_W} \#(w) \cdot k \cdot \frac{\#(c)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c}) \\
&+ \sum_{w \in V_W} \sum_{c_N \in V_C \setminus \{c\}} \#(w) \cdot k \cdot \frac{\#(c_N)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c_N})
\end{aligned}
\tag{7}
$$

This reveals the objective for a *specific* word-context (w,c) pair:

$$
L(w,c) = \#(w,c) \cdot \log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \#(w) \cdot \frac{\#(c)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c}) \tag{8}
$$

Let $x = \vec{w} \cdot \vec{c}$, and take the partial derivative with respect to x:

$$
\frac{\partial L}{\partial x} = \#(w,c) \cdot \sigma(-x) - k \cdot \#(w) \cdot \frac{\#(c)}{|D|} \cdot \sigma(x) \tag{9}
$$

To optimize the objective, we set the derivative to zero [1].

$$
\begin{aligned}
\frac{\partial L}{\partial x} &= \#(w,c) \cdot \frac{1}{1+e^x} - k \cdot \#(w) \cdot \frac{\#(c)}{|D|} \cdot \frac{1}{1+e^{-x}} \\
&= \#(w,c) \cdot \frac{1}{1+e^x} - k \cdot \#(w) \cdot \frac{\#(c)}{|D|} \cdot \frac{e^x}{1+e^x} \\
&= \frac{1}{1+e^x}\Big(\#(w,c) - k \cdot \#(w) \cdot \frac{\#(c)}{|D|} \cdot e^x\Big) = 0 \\
\Rightarrow \#(w,c) &= k \cdot \#(w) \cdot \frac{\#(c)}{|D|} \cdot e^x \Rightarrow e^x = \frac{\#(w,c)}{k \cdot \#(w) \cdot \frac{\#(c)}{|D|}} \\
\Rightarrow x = \vec{w} \cdot \vec{c} &= \log\left(\frac{\#(w,c)}{\#(w) \cdot \frac{\#(c)}{|D|}} \cdot \frac{1}{k}\right) = \log\left(\frac{\#(w,c)|D|}{\#(w) \cdot \#(c)}\right) - \log k
\end{aligned}
\tag{10}
$$

The expression $\log\left(\frac{\#(w,c)|D|}{\#(w)\cdot\#(c)}\right)$ is the well-known point-wise mutual information (PMI) of $\#(w,c)$ used widely in NLP. For negative-sampling value of $k = 1$, the SGNS objective is factorizing a word-context matrix in which the association is measured by $f(w,c) = \text{PMI}(w,c)$. We denote the PMI matrix as M^{PMI}. For negative-sampling values $k > 1$, SGNS is factorizing a *shifted* PMI matrix $M^{PMI_k} = M^{PMI} - \log k$.

2.2 Point-wise Mutual Information

Point-wise mutual information is an information-theoretic association measure between a pair of discrete outcomes x and y, defined as:

$$\text{PMI}(x,y) = \log\frac{P(x,y)}{P(x)P(y)} \tag{11}$$

In our case, $\text{PMI}(w,c)$ can be measured empirically as:

$$\text{PMI}(w,c) = \log\frac{\#(w,c)\cdot|D|}{\#(w)\cdot\#(c)} \tag{12}$$

The matrix is ill-defined for pairs that were never observed in the corpus, i.e. $\text{PMI}(w,c) = \log 0 = -\infty$. A sparse, consistent and common alternative is to use the *positive* PMI (PPMI) metric, in which all negative values are replaced by 0:

$$\text{PPMI}(w,c) = \max(\text{PMI}(w,c), 0) \tag{13}$$

2.3 Weighted Matrix Factorization

The assumption of having perfect reconstruction is not possible; hence, some $\vec{w}\cdot\vec{c}$ products must deviate from their optimal values. The pair-specific objective equation 8 reveals that the loss for a pair (w,c) depends on its number of observations $\#(w,c)$ and expected negative samples $k\cdot\#(w)\cdot\frac{\#(c)}{|D|}$. SGNS's objective can now be cast as a *weighted matrix factorization* problems, seeking the optimal d-dimensional factorization of the matrix $M^{PMI} - \log k$ under a metric which pays more for deviations on frequent (w,c) pairs than deviations on infrequent ones.

An alternative matrix factorization is factorizing the PPMI matrix with truncated SVD. The word and context representations can be obtained by $W^{SVD} = U_d \cdot \Sigma_d$ and $C^{SVD} = V_d$ or $W^{SVD} = U_d \cdot \sqrt{\Sigma_d}$ and $C^{SVD} = V_d \cdot \sqrt{\Sigma_d}$. The symmetric SVD works better empirically although it is not theoretically clear why.

An interesting middle-ground between SGNS and SVD is the use of stochastic matrix factorization (SMF) approaches, common in the collaborative filtering literature. The exploration of SMF-based algorithms for word embeddings is left for future work.

[1] The derivation here is a little bit cleaner (in my opinion) than the one presented in the paper. However, the results are the same.

References

[1] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc., 2014.