Asymmetric Word-Embedding Proposal

Hao Xue

March 22 2020

1 SGNG

The loss function of the skip-gram with negative sampling model (SGNS) is proposed by Levy and Goldberg[1] in formula (3) (see [2] for detailed derivation of SGNS):

$$\begin{aligned} l &= \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c})) + \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c)(k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c})]) \\ &= \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c})) + \sum_{w \in V_W} \#(w)(k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c})]) \end{aligned} \tag{1}$$

where the expectation term is:

$$\begin{aligned} \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c_N})] &= \sum_{c_N \in V_C} \frac{\#(c_N)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c_N}) \\ &= \frac{\#(c)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c}) + \sum_{c_N \in V_C \setminus \{c\}} \frac{\#(c_N)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c_N}) \end{aligned} \tag{2}$$

Combining equations (1) and (2) results in the local objective for a specific (w, c) pair:

$$l(w,c) = \#(w,c) \log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \#(w) \cdot \frac{\#(c)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c}) \tag{3}$$

Equivalently, in our notations:

$$l = -\sum_{i,j} \#(w_i, c_j) \log \sigma(\mathbf{w}_i^T \mathbf{c}_j) - \frac{k}{|D|} \#(w) \#(c) \log \sigma(-\mathbf{w}_i^T \mathbf{c}_j).$$

By taking the partial derivative with respect to $\mathbf{w}_i$, we get:

$$\frac{\partial}{\partial \mathbf{w}_i} l = -\sum_j \#(w_i, c_j) \sigma(-\mathbf{w}_i^T \mathbf{c}_j) \mathbf{c}_j^T + \frac{k}{|D|} \#(w_i) \#(c_j) \log \sigma(-\mathbf{w}_i^T \mathbf{c}_j) \mathbf{c}_j^T.$$

(But in your slide, it is:

$$\frac{\partial}{\partial \mathbf{w}_i} l = \sum_j (\#(w_i, c_j) + \frac{k}{|D|} \#(w_i) \#(c_j))(\sigma(P_{ij}) - \sigma(\mathbf{w}_i^T \mathbf{c}_j)),$$

where $P_{ij} = PMI(w_i, c_j)$. I guess it is a typo.)

To optimize the objective, the authors define $x = \vec{w} \cdot \vec{c}$ and take the partial order derivative with respect to x (instead of $\mathbf{w}_i$) in (3):

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

By comparing the derivative to zero, they get:

$$e^{2x} - (\frac{\#(w,c)}{k \cdot \#(w) \cdot \frac{\#(c)}{|D|}} - 1)e^x - \frac{\#(w,c)}{k \cdot \#(w) \cdot \frac{\#(c)}{|D|}} = 0,$$

which is a quadratic equation of e^x, with the only feasible solution:

$$e^x = \frac{\#(w,c) \cdot |D|}{\#(w)\#(c)} \cdot \frac{1}{k}.$$

Substituting x with $\vec{w} \cdot \vec{c}$ reveals:

$$\begin{aligned}\vec{w} \cdot \vec{c} &= \log(\frac{\#(w,c) \cdot |D|}{\#(w)\#(c)} \cdot \frac{1}{k}) \\ &= \log(\frac{\#(w,c) \cdot |D|}{\#(w)\#(c)}) - \log k.\end{aligned}$$

Therefore, they conclude that SGNS can be cast into factorizing a shifted positive PMI matrix, say P, whose i-th row j-th column is $P_{i,j} = \max(\log(\frac{\#(w_i,c_j) \cdot |D|}{\#(w_i)\#(c_j)}) - \log k, 0)$. Again, with our notation the optimal solution to the loss is:

$$\mathbf{w}_i^T \mathbf{c}_j = \log(\frac{\#(w_i,c_j) \cdot |D|}{\#(w_i)\#(c_j)}) - \log k. \tag{4}$$

Then the embedding of word and context could be found by SVD. Let $P = U\Sigma V^T$, Σ_d be the diagonal matrix formed from the top d singular values and let U_d and V_d be the matrices produced by selecting the corresponding columns from U and V. According to [1], the embedding of word and context are respectively:

$$W = U_d\sqrt{\Sigma_d}, \quad C = V_d\sqrt{\Sigma_d}$$

2 Directional SGNG

We can modify the loss into the directional version:

$$\begin{aligned}l = -\sum_i \sum_j \big\{&N(c_j \to w_j) \log \sigma(\mathbf{w}_i^T \mathbf{c}_j^-) + \frac{k}{D^-} N(w_i^-) N(c_j^-) \log \sigma(-\mathbf{w}_i^T \mathbf{c}_j^-) \\ &+ N(w_i \to c_j) \log \sigma(\mathbf{w}_i^T \mathbf{c}_j^+) + \frac{k}{D^+} N(w_i^+) N(c_j^+) \log \sigma(-\mathbf{w}_i^T \mathbf{c}_j)\big\}\end{aligned} \tag{5}$$

(the sign before $\frac{k}{D^-}$ should be positive), where

$$\begin{aligned}N(w_i^+) = \sum_j N(w_i \to c_j), \quad N(w_i^-) = \sum_j N(c_j \to w_i), \\ N(c_j^+) = \sum_i N(w_i \to c_j), \quad N(c_j^-) = \sum_i N(c_j \to w_i),\end{aligned}$$

with local loss for a particular pair $(\mathbf{w}_i, \mathbf{c}_j)$:

$$\begin{aligned}l(\mathbf{w}_i, \mathbf{c}_j) = -(&N(c_j \to w_j) \log \sigma(\mathbf{w}_i^T \mathbf{c}_j^-) + \frac{k}{D^-} N(w_i^-) N(c_j^-) \log \sigma(-\mathbf{w}_i^T \mathbf{c}_j^-) \\ &+ N(w_i \to c_j) \log \sigma(\mathbf{w}_i^T \mathbf{c}_j^+) + \frac{k}{D^+} N(w_i^+) N(c_j^+) \log \sigma(-\mathbf{w}_i^T \mathbf{c}_j))\end{aligned} \tag{6}$$

we can get the partial derivative of (5) with respect to $\mathbf{w}_i$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}_i} l = -\sum_j \big\{&N(c_j \to w_i) \sigma(-\mathbf{w}_i^T \mathbf{c}_j^-)(\mathbf{c}_j^-)^T - \frac{k}{D^-} N(w_i^-) N(c_j^-) \sigma(\mathbf{w}_i^T \mathbf{c}_j^-)(\mathbf{c}_j^-)^T \\ &+ N(w_i \to c_j) \sigma(-\mathbf{w}_i^T \mathbf{c}_j^+)(\mathbf{c}_j^+)^T - \frac{k}{D^+} N(w_i^+) N(c_j^+) \sigma(\mathbf{w}_i^T \mathbf{c}_j^+)(\mathbf{c}_j^+)^T \big\}.\end{aligned}$$

If we want to mimic the derivation of optimal solution in [1], we need to define $u = \mathbf{w}_i^T \mathbf{c}_j^-$ and $v = \mathbf{w}_i^T \mathbf{c}_j^+$. Then, taking the derivative of 6 with respect to u and v separately,

$$\frac{\partial l}{\partial u} = -N(c_j \to w_i)\sigma(-u) + \frac{k}{D^-} N(w_i^-)N(c_j^-)\sigma(u)$$
$$\frac{\partial l}{\partial v} = -N(w_i \to c_j)\sigma(-v) + \frac{k}{D^+} N(w_i^+)N(c_j^+)\sigma(v).$$

Therefore, by following the same way we get (4), we have:

$$\mathbf{w}_i^T \mathbf{c}_j^- = \log(\frac{N(c_j \to w_i) \cdot |D|}{N(w_i^-)N(c_j^-)}) - \log k$$
$$\mathbf{w}_i^T \mathbf{c}_j^+ = \log(\frac{N(w_i \to c_j) \cdot |D|}{N(w_i^+)N(c_j^+)}) - \log k.$$

Similarly, we can get the optimal solution by factorizing $P^- - \log k$ and $P^+ - \log k$, where P^- is the left shifted PPMI matrix with element, $P_{ij}^- = \max(\log(\frac{N(c_j \to w_i) \cdot |D|}{N(w_i^-)N(c_j^-)}) - \log k, 0)$, and P^+ is the right sifted PPMI matrix with element, $P_{ij}^+ = \max(\log(\frac{N(w_i \to c_j) \cdot |D|}{N(w_i^+)N(c_j^+)}) - \log k, 0)$. But we will reach a problem if we follow this method, that is, we cannot ensure that the left component of the two matrix factorization to be the same. To be more specific, if the SVD of two matrices are

$$P^- = U^- \Sigma^- (V^-)^T$$
$$P^+ = U^+ \Sigma^+ (V^+)^T,$$

the word and context embeddings are:

$$\begin{aligned} W &= U_d^+ \sqrt{\Sigma_d^+} \\ &= U_d^- \sqrt{\Sigma_d^-} \\ C^+ &= V_d^+ \sqrt{\Sigma_d^+} \\ C^- &= V_d^- \sqrt{\Sigma_d^-} \end{aligned} \tag{7}$$

however, the equality of (8) is not guaranteed to hold.

3 Alternating Minimization

To solve the optimization problem:

$$\min \|WC^+ - P^+\|_F^2 + \|WC^- - P^-\|_F^2$$

Alternation Minimization could be used, which iterates solving the three sub-problems below:

$$W_{(t+1)} = \arg\max_{\mathrm{W}} \|WC_{(t)}^+ - y\|_F^2 + \|WC_{(t)}^- - z\|_F^2 \tag{8}$$
$$C_{(t+1)}^+ = \arg\max_{\mathrm{C}^+} \|W_{(t+1)}C^+ - y\|_F^2 \tag{9}$$
$$C_{(t+1)}^- = \arg\max_{\mathrm{C}^-} \|W_{(t+1)}C^- - z\|_F^2 \tag{10}$$

We can use the updating formulae in this note for the i-th element of the j-th column $(C_{ij}^+)^{(t+1)}$ and $(C_{ij}^-)^{(t+1)}$ to solve (9) and (10):

$$(C_{ij}^+)^{(t+1)} = (C_{ij}^+)^{(t)} \frac{[W^T P_{\cdot j}^+]_i}{[W^T W (C_{\cdot j}^+)^{(t)}]_i}$$
$$(C_{ij}^-)^{(t+1)} = (C_{ij}^-)^{(t)} \frac{[W^T P_{\cdot j}^-]_i}{[W^T W (C_{\cdot j}^-)^{(t)}]_i},$$

where y and z are respectively, the column vector of y and z. To solve (8), we can use a similar method. Now, consider $F(W) = \|WC^+ - P^+\|_F^2 + \|WC^- - P^-\|_F^2 = \|(C^+)^T W^T - (P^+)^T\|_F^2 + \|(C^-)^T W^T - (P^-)^T\|_F^2$ with C^+ and C^- fixed. Since $F(W)$ is separable over columns of W^T, we define $f(w) = \frac{1}{2}\|y - Aw\|_2^2 + \frac{1}{2}\|z - Bw\|_2^2$. We can construct an auxiliary function $g(w, \tilde{w})$ used in the note:

$$\begin{aligned} g(w,\tilde{w}) = &\frac{1}{2}\|y\|_2^2 - \sum_i y_i (A_i)^T w + \frac{1}{2}\sum_i \sum_j \lambda_{ij} (\frac{A_{ij} w_j}{\lambda_{ij}})^2 \\ &+ \frac{1}{2}\|z\|_2^2 - \sum_i z_i (B_i)^T w + \frac{1}{2}\sum_i \sum_j \mu_{ij} (\frac{B_{ij} w_j}{\mu_{ij}})^2, \end{aligned}$$

where $\lambda_{ij} = \frac{A_{ij}\tilde{w}_j}{\sum_k C_{ik}^+ \tilde{w}_k} = \frac{A_{ij}\tilde{w}_j}{(A_i)^T \tilde{w}}$ and $\mu_{ij} = \frac{B_{ij}\tilde{w}_j}{\sum_k B_{ik}\tilde{w}_k} = \frac{B_{ij}\tilde{w}_j}{(B_i)^T \tilde{w}}$. Then we have:

$$g(w, w) = f(w),$$

and

$$g(w, \tilde{w}) \geq f(w)$$

since

$$\begin{aligned} f(w) &= \frac{1}{2}\|y - Aw\|_2^2 + \|z - Bw\|_2^2 \\ &= \frac{1}{2}\|y\|_2^2 - \sum_i y_i (A_i)^T w + \frac{1}{2}\sum_i \sum_j (\lambda_{ij} \frac{A_{ij} w_j}{\lambda_{ij}})^2 \\ &+ \frac{1}{2}\|z\|_2^2 - \sum_i z_i (B_i)^T w + \frac{1}{2}\sum_i \sum_j (\mu_{ij} \frac{B_{ij} w_j}{\mu_{ij}})^2 \\ &\leq \frac{1}{2}\|y\|_2^2 - \sum_i y_i (A_i)^T w + \frac{1}{2}\sum_i \sum_j \lambda_{ij} (\frac{A_{ij} w_j}{\lambda_{ij}})^2 \\ &+ \frac{1}{2}\|z\|_2^2 - \sum_i z_i (B_i)^T w + \frac{1}{2}\sum_i \sum_j \mu_{ij} (\frac{B_{ij} w_j}{\mu_{ij}})^2 \\ &= g(w, \tilde{w}) \end{aligned}$$

Therefore we have the following descent by denoting $w^{(t+1)} = \arg\max_{\mathrm{w}} g(w, w^{(t)})$:

$$f(w^{(t+1)}) \leq g(w^{(t+1)}, w^{(t)}) \leq g(w^{(t)}, w^{(t)}) = f(w^{(t)}, w^{(t)}).$$

We can find each element of $w^{(t+1)}$, say $w_k^{(t+1)}$, by setting $\frac{\partial}{\partial w_k} g(w, w^{(t)})$ into zero:

$$w_k^{(t+1)} = \frac{\sum_i y_i A_{ik} + \sum_i z_i B_{ik}}{\sum_i c_{ik}^+ c_i^+ w^{(t)} + \sum_i c_{ik}^- c_i^- w^{(t)}} w_k^{(t)} = \frac{[A^T y + B^T z]_k}{[A^T A w_k^{(t)} + B^T B w_k^{(t)}]_k} w_k^{(t)}. \tag{11}$$

Substitute $A = (C^+)^T$, $B = (C^-)^T$, $y = (P_{i.}^+)^T$, $z = (P_{i.}^-)^T$ and $w = w_{i.}$ into (11), we get

$$w_{ij}^{(t+1)} = \frac{[C^+ (P_{i.}^+)^T + C^- (P_{i.}^-)^T]_j}{[C^+ (C^+)^T (w_{i.}^{(t)})^T + [C^- (C^-)^T (w_{i.}^{(t)})^T]_j} w_{ij}^{(t)}$$

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.

[2] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, 2014.