A Critique of the Smooth Inverse Frequency Sentence Embeddings (Student Abstract)

Aidana Karipbayeva
College of Liberal Arts and Sciences
University of Illinois at Urbana-Champaign
aidana2@illinois.edu

Alena Sorokina
School of Sciences and Humanities
Nazarbayev University
alena.sorokina@nu.edu.kz

Zhenisbek Assylbekov
School of Sciences and Humanities
Nazarbayev University
zhassylbekov@nu.edu.kz

Abstract

We critically review the smooth inverse frequency sentence embedding method of Arora, Liang, and Ma (2017), and show inconsistencies in its setup, derivation and evaluation.

Introduction

The smooth inverse frequency (SIF) sentence embedding method of Arora, Liang, and Ma (2017) has gained attention in the NLP community due to its simplicity and competetive performance. We recognize the strengths of this method, but we argue that its theoretical justification contains a number of flaws. In what follows we show that there are contradictory arguments in the setup, derivation and experimental evaluation of SIF.

We first recall the word production model used by the authors as the foundation of SIF: given the context vector $\mathbf{c} \in \mathbb{R}^d$, the probability that a word w is emitted in the context is modeled by

$$p(w \mid \mathbf{c}) = \alpha p(w) + (1-\alpha)\exp(\langle \mathbf{w}, \tilde{\mathbf{c}}\rangle)/Z_{\tilde{c}} \qquad (1)$$

$$\text{with } \tilde{\mathbf{c}} = \beta \mathbf{c}_0 + (1-\beta)\mathbf{c}, \quad \mathbf{c}_0 \perp \mathbf{c}, \qquad (2)$$

where $\alpha, \beta \in [0,1]$ are scalar hyperparameters, $\mathbf{w} \in \mathbb{R}^d$ is a word embedding for w, $\mathbf{c}_0 \in \mathbb{R}^d$ is the so-called common discourse, and $Z_{\tilde{\mathbf{c}}} = \sum_{w \in \mathcal{W}} \exp(\langle \tilde{\mathbf{c}}, \mathbf{w}\rangle)$ is the normalizing constant.

Inconsistent Setup

The authors empirically find (see their section 4.1.1) that the optimal value of α satisfies

$$10^{-4} \le (1-\alpha)/(\alpha Z) \le 10^{-3}, \qquad (3)$$

where $Z = \mathbb{E}[Z_\mathbf{c}]$. In their previous work, Arora et al. (2016) showed (see the proof sketch of their Lemma 2.1 on p. 398) that under isotropic assumption on $\mathbf{w}$'s,

$$\mathbb{E}_\mathbf{w}[Z_\mathbf{c}] = n\mathbb{E}_\xi[\exp(\xi^2\|\mathbf{c}\|^2/2)] \qquad (4)$$

where ξ is a random variable upper bounded by a constant. From (4) we have $n \le Z = \mathbb{E}[Z_\mathbf{c}]$, and combining this with the right inequality from (3) we have $\frac{1-\alpha}{\alpha} \le \frac{1}{10^3 n} \Leftrightarrow \frac{10^3 n}{10^3 n+1} \le \alpha$. For a typical vocabulary size $n = 10^5$ this implies $0.99999999 \le \alpha \le 1$, which means that the generative model (1) is essentially a unigram model $\Pr(w \mid \mathbf{c}) \approx p(w)$ that practically ignores the context.

Contradictory Derivation

Treating any sentence s as a sequence of words $[w]$, the authors construct its log-likelihood given the smoothed context vector $\tilde{\mathbf{c}}$ as $\ell(\mathbf{c}) = \sum_{w \in s} \log \Pr(w \mid \mathbf{c})$. Then this log-likelihood is linearized using Taylor expansion at $\tilde{\mathbf{c}} = \mathbf{0}$:

$$\ell(\tilde{\mathbf{c}}) \approx \ell(\mathbf{0}) + \nabla\ell(\mathbf{0})^\top \tilde{\mathbf{c}}, \qquad (5)$$

and after that the right-hand side of (5) is optimized with $\tilde{\mathbf{c}}$ constrained to take values on the unit sphere $\{\tilde{\mathbf{c}} \in \mathbb{R}^d \mid \|\tilde{\mathbf{c}}\| = 1\}$, which contradicts the assumption $\tilde{\mathbf{c}} \approx \mathbf{0}$ needed for the linear approximation (5) to be adequate.

Model Inadequacy

It *is* possible to have a valid derivation of the SIF sentence embedding as the Maximum-a-Posteriori (MAP) estimate of $\mathbf{c}$ given s once we assume the generative model

$$p(w \mid \mathbf{c}) \propto \exp\left(\tfrac{a}{p_i+a}\langle \mathbf{w}, \tilde{\mathbf{c}}\rangle\right) \qquad (6)$$

instead of (1). The proof is similar to that of Lemma 3.1 in Arora et al. (2016). Now, assume that c is a single context word, and $\mathbf{c}$ is its embedding. Taking logarithm of both sides in (6), then solving for $\langle \mathbf{w}, \tilde{\mathbf{c}}\rangle$ and assuming that the normalizer in (6) concentrates well around a constant Z, we have

$$\langle \mathbf{w}, \tilde{\mathbf{c}}\rangle \approx \tfrac{p(w)+a}{a}\left(\log p(w \mid c) + \log Z\right) \qquad (7)$$

This means that the word and context embeddings that underlie the language model (6) give a low-rank approximation of a matrix $\mathbf{M}$ in which the element in row w and column c is equal to the right-hand side of (7). It is well known that the word and context embeddings that underlie the SGNS training objective give a low-rank approximation of the shifted PMI matrix, and that factorizing the latter with truncated SVD gives embeddings of similar quality (Levy and Goldberg 2014). This means, that if the model (6) is adequate,

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

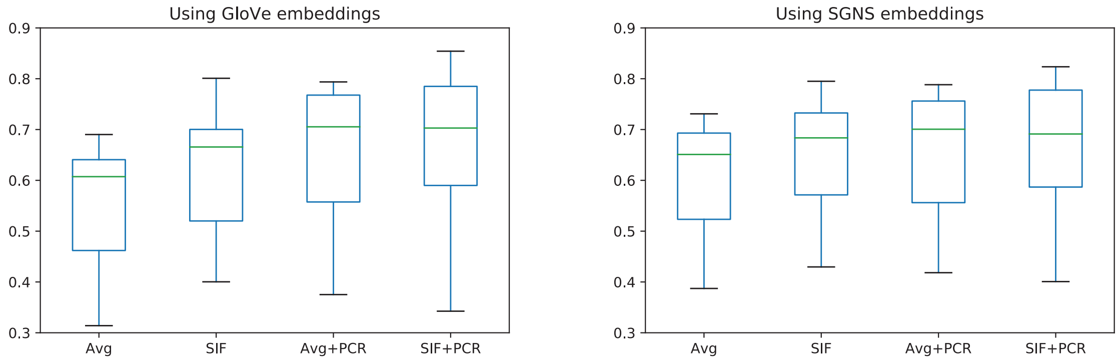

Figure 1: Performance of SIF vs Simple Average on STS tasks. The evaluation criterion is the Pearson's coefficient between the predicted scores and the ground-truth scores.

Matrix	Similarity task	Analogy Task
$\mathbf{M}$	61.18	27.14
PMI	65.05	29.58

Table 1: Word embeddings from PMI and $\mathbf{M}$. For word similarities evaluation metric is the Spearman's correlation with the human ratings, while for word analogies it is the percentage of correct answers.

then the truncated SVD of $\mathbf{M}$ should give us good-quality word embeddings as well. We calculated the shifted PMI and $\mathbf{M}$ on `text8` data[1] using vocabulary size 35000 and then performed rank-200 approximation. The resulting embeddings were evaluated on standard similarity (WordSim) and analogy (Google and MSR) tasks. The hyperparameter Z was tuned using grid search and the optimal value was s.t. $\log Z = 13$. The results of evaluation are given in Table 1. As we can see the word embeddings that underlie (6) do not outperform those that underlie SGNS. Hence, the adequacy of (6) is questionable.

Incomplete Evaluation

The method of Arora, Liang, and Ma (2017) is not only in using SIF weights but also in removing the principal component from the resulting sentence embeddings. When the authors evaluate their method against a simple average (Avg) of word vectors, they do not consider principal component removal (PCR) as a separate factor, i.e. they do not compare against a simple average of word embeddings followed by a principal component removal (Avg+PCR). We performed such comparison on datasets from the SemEval Semantic Textual Similarity (STS) tasks[2] with GLOVE and SGNS embeddings[3], and the results are illustrated on Fig. 1. As we can see, SIF is indeed stronger than Avg, but this advantage is diminished when we remove the principal components from both. Looking at the boxplots, one may think that the difference between Avg+PCR and SIF+PCR is not significant, however this is not the case: SIF+PCR demonstrates higher scores than Avg+PCR according to paired one-sided Wilcoxon signed rank test, with p-values < 0.02 for both GloVe and SGNS embeddings. Thus, we admit that the overall claim of the authors is valid: SIF outperforms Avg with and without PCR.

Conclusion

The sentence embedding method of Arora, Liang, and Ma (2017) is indeed a simple but tough-to-beat baseline, which has a clear underlying intuition that the embeddings of too frequent words should be downweighted when summed with those of less frequent ones. However, one does not need to tweak a previously developped mathematical theory to justify this empirical finding: in pursuit of mathematical validity, the SIF authors made their theoretical argument doubtful in a number of ways.

Acknowledgements

This work is supported by the Nazarbayev University Collaborative Research Program 091019CRP2109, and by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan, IRN AP05133700.

References

Arora, S.; Li, Y.; Liang, Y.; Ma, T.; and Risteski, A. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics* 4:385–399.

Arora, S.; Liang, Y.; and Ma, T. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR*.

Levy, O., and Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, 2177–2185.

[1] http://mattmahoney.net/dc/textdata.html

[2] http://ixa2.si.ehu.es/stswiki/index.php/Main_Page

[3] Our code is available at https://github.com/the-regressionists/Critique-SIF-Sent-Embeddings