Less is More: Filtering Abnormal Dimensions in GloVe

Yang-Yin Lee, Hao Ke, Hen-Hsen Huang, Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan
{yylee, hke, hhhuang}@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

ABSTRACT

GloVe, global vectors for word representation, performs well in some word analogy and semantic relatedness tasks. However, we find that some dimensions of the trained word embedding are abnormal. We verify our conjecture via removing these abnormal dimensions using Kolmogorov–Smimov test and experiment on several benchmark datasets for semantic relatedness measurement. The experimental results confirm our finding. Interestingly, some of the tasks outperform the state-of-the-art model SensEmbed by simply removing these abnormal dimensions. The novel rule of thumb technique which leads to better performance is expected to be useful in practice.

Keywords

GloVe; Semantic relatedness; word embedding;

1. INTRODUCTION

GloVe [6], a log-bilinear regression model proposed recently, tries to resolve the drawbacks of the global factorization approaches (e.g., latent semantic analysis [2]) and the local context window approaches (e.g., skip-gram model [5]) on the word analogy and the semantic relatedness task. The global vectors in GloVe are trained using unsupervised learning on aggregated global word-word co-occurrence statistics from a corpus. Consider an example "*solid* is more related to *ice* and *gas* is more related to *steam*" to show the idea behind. GloVe let the ratio of the probability $P(k|ice)/P(k|steam)$ be high if $k = solid$ and low if $k = gas$. If $k = water$ or $k = fashion$, then the ratio should close to 1 because both *ice* and *steam* are equally related to *water* and equally unrelated to *fashion*. The probability can be derived from the co-occurrence matrix, and GloVe utilizes this ratio of probability to capture the relatedness between words.

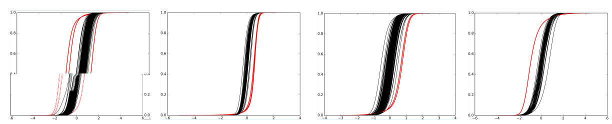
The objective of GloVe is to factorize the log-count matrix and to find the word embedding that satisfies this ratio. However, we find that some dimensions are abnormal in every trained word embedding. We suspect that the parameters in GloVe are not tuned to globally optimized values. In this paper, we explore the Kolmogorov–Smimov test of normality to identify and remove these dimensions and perform the semantic relatedness task to confirm our finding. The experimental results show that a large performance gain can be obtained when these abnormal dimensions are removed.

2. OUR APPROACH

The abnormal dimensions in each word embedding model are selected in the following way: (a) for each dimension, compute the Kolmogorov–Smimov test statistic, (b) sort the test statistic in descending order, (c) select the dimensions with the statistic values greater than 41, and (d) if no statistic value is greater than 41 in the word embedding, then select the top two dimensions.

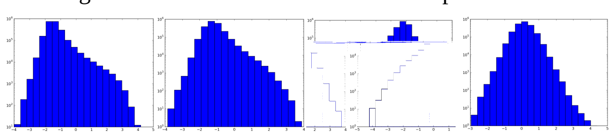
We explore three versions of the GloVe pre-trained word vectors[1]: (1) 6B tokens, 400K vocab, uncased, 50d, 100d, 200d and 300d vectors trained on the Wikipedia 2014 and Gigaword 5, (2) 42B tokens, 1.9M vocab, uncased, 300d vectors trained on the Common Crawl, and (3) 840B tokens, 2.2M vocab, cased, 300d vectors trained on the Common Crawl.

Figure 1 shows the empirical CDF of four GloVe embeddings. In each subplot, black lines are the normal dimensions while red lines are the abnormal dimensions. Figure 2 shows the shapes of the three abnormal dimensions, i.e., dim 10, 18 and 141, removed from GloVe 840B 300d by the aforementioned algorithm. The new model is called GloVe 840B 297d. The dimensions removed from the other versions are listed as follows: (1) GloVe 6B 49d (remove dim 31), (2) GloVe 6B 98d (remove dim 56 and 59), (3) GloVe 6B 199d (remove dim 22), (4) GloVe 6B 298d (remove dim 277 and 10), and (5) GloVe 42B 297d (remove dim 225, 7 and 97).



(a) 840B 300d (b) 42B 300d (c) 6B 300d (d) 6B 50d

Figure 1. Different GloVe versions' Empirical CDF



(a) dim 10 (b) dim 18 (c) dim 141 (d) dim 150

Figure 2. Abnormal dimensions (a-c) and normal dimension (d)

3. EXPERIMENTS

3.1 Datasets

We use cosine similarity to compute the semantic relatedness of a pair of words represented by different versions of GloVe. Six benchmark datasets are considered in the experiments: RG-65 [7], WordSim353 (WS353-sim, WS353-rel) [3], YP130 [8], and MEN [1]. The RG-65 word similarity dataset consists of 65 word pairs. For each word pair, there is a rating score, ranging from 0.0 to 4.0 to denote *semantically unrelated* to *highly synonymous* by 51 subjects. WordSim353 (WS353-all) contains 353 word pairs whose scores range from 0.0 to 10.0. It includes two subsets for measuring similarity (WS353-sim) and relatedness (WS353-rel), respectively. The YP130 dataset is designed specifically for measuring the verb similarity. The MEN dataset is composed of two sets of English word pairs with human-assigned similarity

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
DOI: http://dx.doi.org/10.1145/2872518.2889381.

[1] http://nlp.stanford.edu/projects/glove/

Table 1. Spearman (ρ) and Pearson (*r*) correlation of different semantic relatedness measures on RG-65, WS353-all, WS353-sim, WS353-rel, YP130 and MEN datasets.

Method	RG-65	WS353-all	WS353-sim	WS353-rel	YP130	MEN
	”/r	”/r	”/r	”/r	”/r	”/r
SensEmbed	**0.894**/None	0.714/None	0.756/None	0.645/None	**0.734**/None	0.779/None
SensEmbed	0.871/None	**0.779**/None	**0.812**/None	0.703/None	0.639/None	0.805/None
Word2Vec	0.761/0.772	0.694/0.649	0.777/0.768	0.622/0.583	0.570/0.589	0.782/0.770
GloVe 6B 50d	0.595/0.557	0.503/0.507	0.573/0.545	0.466/0.503	0.400/0.374	0.657/0.667
GloVe 6B 49d	0.600/0.592	0.582/0.586	0.646/0.645	0.546/0.559	0.427/0.423	0.690/0.696
GloVe 6B 100d	0.676/0.674	0.533/0.548	0.604/0.594	0.496/0.543	0.475/0.470	0.693/0.697
GloVe 6B 98d	0.686/0.693	0.607/0.617	0.673/0.673	0.588/0.606	0.507/0.516	0.723/0.723
GloVe 6B 200d	0.713/0.717	0.578/0.578	0.629/0.625	0.545/0.563	0.537/0.543	0.724/0.725
GloVe 6B 199d	0.731/0.733	0.623/0.617	0.679/0.686	0.580/0.584	0.566/0.578	0.752/0.749
GloVe 6B 300d	0.770/0.752	0.609/0.604	0.664/0.665	0.573/0.580	0.580/0.583	0.749/0.743
GloVe 6B 298d	0.766/0.756	0.653/0.637	0.706/0.712	0.608/0.599	0.599/0.610	0.769/0.761
GloVe 42B 300d	0.817/0.800	0.632/0.639	0.698/0.704	0.571/0.603	0.502/0.467	0.744/0.742
GloVe 42B 297d	0.811/**0.822**	0.773/**0.733**	0.797/0.793	**0.745**/**0.718**	0.590/**0.612**	0.811/0.804
GloVe 840B 300d	0.770/0.771	0.712/0.710	0.802/0.803	0.644/0.658	0.540/0.522	0.805/0.807
GloVe 840B 297d	0.759/0.761	0.764/0.737	0.806/**0.813**	0.716/0.703	0.566/0.585	**0.827**/**0.829**

scores. The comparison models are the state-of-the-art approach SensEmbed [4] and Word2Vec [5].

3.2 Results and Discussion

Table 1 shows the Spearman (ρ) and Pearson (r) correlation of different semantic relatedness measures on RG-65, WS353-all, WS353-sim, WS353-rel, YP130, and MEN datasets. The dimension-removed versions of the GloVe model are listed below the origin versions. Comparing to the approaches without removing (e.g., GloVe 6B 50d vs. GloVe 6B 49d, and GloVe 840B 300d vs. GloVe 840B 297d), we can find that the removal of the abnormal dimensions is indeed beneficial to the semantic relatedness tasks under different sizes of training corpus and dimensionality. For the GloVe 6B models without/with removal, the performance is directly proportional to the dimensionality. That is in accord with the original GloVe paper's findings. The 840B version is not always better than the 42B version. The possible reason may be that 840B version is case-sensitive. Interestingly, the ” of the GloVe 42B gain from 0.571 to 0.745, and from 0.744 to 0.811 in the WS353-rel and MEN datasets, respectively, after removing abnormal dimensions.

The two abnormal dimension removal approaches, GloVe 42B 297d and GloVe 840B 297d, outperform the state-of-the-art SensEmbed's approach on WS353-rel and MEN datasets. The performance of all the other models is still lower than that of SensEmbed. The reason may be that GloVe does not disambiguate each word's senses during its training phase, and that is the main contribution in the SensEmbed's research.

4. CONCLUSIONS

In this paper we show that the GloVe model produces some abnormal dimensions. The Kolmogorov–Smimov test of normality is applied to determine those dimensions. The experimental results show that the removal of the abnormal dimensions is indeed beneficial to the trained vectors for word relatedness measurement. The GloVe model with the abnormal dimension removal outperforms one of the state-of-the-art method SensEmbed in two benchmark datasets. In the end, we would like to address some related issues: (1) Can we avoid producing those abnormal dimensions during the training phase of the GloVe? (2) Besides directly removing those abnormal dimensions, are there any ways to refine or correct those dimensions? (3) What is the critical procedure in the GloVe that produces those abnormal dimensions? (4) What is the physical meaning behind the abnormal dimensions?

5. ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-102-2221-E-002-103-MY3 and MOST-104-2221-E-002-061-MY3.

6. REFERENCES

[1] Bruni, E. et al. 2014. Multimodal Distributional Semantics. *J. Artif. Intell. Res.(JAIR)*. 49, (2014), 1–47.

[2] Deerwester, S.C. et al. 1990. Indexing by latent semantic analysis. (1990).

[3] Finkelstein, L. et al. 2001. Placing search in context: The concept revisited. *Proceedings of the 10th international conference on World Wide Web* (2001), 406–414.

[4] Iacobacci, I. et al. 2015. SensEmbed: learning sense embeddings for word and relational similarity. *Proceedings of ACL* (2015), 95–105.

[5] Mikolov, T. et al. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* (2013), 3111–3119.

[6] Pennington, J. et al. 2014. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing* (2014), 1532–1543.

[7] Rubenstein, H. and Goodenough, J.B. 1965. Contextual correlates of synonymy. *Communications of the ACM*. 8, 10 (1965), 627–633.

[8] Yang, D. and Powers, D.M. 2005. Measuring semantic similarity in the taxonomy of WordNet. *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38* (2005), 315–322.