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 be high if solid is more related to ice and low if solid is more related 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 using 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 vectors1: (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).

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

1 http://nlp.stanford.edu/projects/glove/
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 (\( \rho \)) 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. For the GloVe 42B model, the removal of the abnormal dimensions is also beneficial. 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–Smirnov 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