UNIVERSITY OF CAMBRIDGE

A Source-Criticism Debiasing Method for GloVe Embeddings

Hope McGovern[1]
[1]Cambridge Computer Laboratory, University of Cambridge

Abstract

We present a simple yet effective method for debiasing GloVe word embeddings [1] which works by incorporating explicit information about training set bias rather than removing biased data outright. We use a fast bias gradient approximation method [2] and borrow the notion of 'source criticism' from the humanities to create our method, Source-Critical GloVe (SC-GloVe). SC-GloVe reduces the effect size on Word Embedding Association Test [3] (WEAT) sets without sacrificing training data or TOP-1 performance.

Bias in Word Embeddings

Word embeddings are numerical representations of text that capture semantic relationships between words. However, word embeddings trained on large public corpora consistently exhibit known human social biases which can be identified with a simple analogy test as presented by Bolukbasi et al. [4]:

Man is to Computer Programmer as woman is to _____

Homemaker
Nurse
Receptionist

To measure biases, we use the Word Embedding Association Test (WEAT), which gives the probability that the observed similarity scores could have arisen with no semantic association between the target concepts and the attribute.

WEAT	Target Sets	Attribute Sets
I	Science/Arts	Male/Female
II	Instruments/ Weapons	Pleasant/ Unpleasant

We consider the **effect size** of two different WEAT bias word sets, one of which reflects a problematic bias and one more benign. The WEAT test measures the similarity of words *a* and *b* in word embedding *w* as measured by the cosine similarity of their vectors, $cos(w_a, w_b)$.

We train an initial GloVe embedding model on a corpus constructed from a Simple English Wikipedia dump[1] using 75-dimensional word embedding vectors. The TOP-1 analogies test measured ~35%.

[1]https://dumps.wikimedia.org/simplewiki

Wiki	
Corpus	
Min. doc. length	200
Max. doc. length	10,000
Num. documents	29,344
Num. tokens	17,033,637
Vocabulary	
Token min. count	20
Vocabulary size	44,806
GloVe	
Context window	symmetric
window size	8
α	0.75
x_{max}	100
Vector Dimension	75
Training epochs	300
Performance	
TOP-1 Analogy	35%

Table 2. Experimental setup for Wiki corpus.

Differential Bias Approximation

Influence functions (IF), a method borrowed from robust statistics, make it possible to determine which inputs to a model exert the most influence over model inference at test time [5]. IFs are an approximation of the result that would be achieved by removing one example at a time from the dataset and training a model to see its net effect on inference.

These tools **provide insight into model behaviour by directly tracking how single training examples affect downstream inference** and therefore hold potential for the growing trend of explainable machine learning [6].

Brunet et al. [2] provide a method to approximate how the optimal word vector learned from the initial training, w_i^*, will change w.r.t. a given corpus perturbation (removing a document):

$$\tilde{w_i} \approx w_i^* - \frac{1}{V} H_{w_i}^{-1} [\nabla_{w_i} L(\tilde{X_i}, w) - \nabla_{w_i} L(X_i, w)] \qquad (2)$$

Where $\tilde{w_i}$ is the word vector learned from a perturbed corpus, V is the size of the vocabulary, X is the global co-occurrence matrix, and X is the co-occurrence matrix only word discounting the co-occurrence matrix of document i. H is the Hessian w.r.t. vector w_i of the point-wise loss at X_i, and $\nabla_{wi} L(X_i, w)$ is the gradient of the point-wise loss function at X_i w.r.t. only word vector w_i.

SC-GloVe

There are two steps to our debiasing method:

- With a single pass through the corpus, use a method of approximating differential bias to generate a weighting factor, β, for each document which corresponds to how much it affects downstream bias at test time.
- Use these weighting factors to update the word vectors relevant to our bias metric to what they *would have been* had the biased document had been counted as 'less reliable' during training. This happens according to:

Algorithm 1 Source-Criticism Debiasing

input *Co-occ Matrix: X, WEAT words: {S,T, A, B}*
  Diff Bias Vector: β,
  $w^*, u^*, b^*, c^* c = GloVe(X)$
 for doc **in** corpus **do**
  # weight the co-occ matrix
  $X' = X - (\beta^{(k)} \cdot X^{(k)})$
  for word i **in** doc $\cap(S \cup T \cup A \cup B)$ **do**
   # approximate WEAT word vecs
   $\tilde{w_i}$ = *# see Equation 2*
   # update WEAT word vecs
   $w_i^* = \tilde{w_i}$
  end for
 end for
re-evaluate WEAT with new word vectors

SC-GloVe (Cont.)

According to Algorithm 1, if the document:

- does NOT affect downstream bias with respect to some bias metric → the weighting factor is zero and the true co-occurrence matrix is used
- INCREASES bias at test time → the co-occurrence matrix is slightly decreased for the relevant words, scaled by how heavily biased it is.
- DECREASES bias downstream → the co-occurrence matrix is slightly increased for the relevant words, strengthening the co-occurrences of more balanced terms.

This yields promising results:

Model	WEAT1	WEAT2
Baseline	0.577	1.04
SC-GloVe	**0.461**	**0.964**

Table 3. WEAT Effect Sizes.

SC-GloVe does decrease the effect size (averaged over 10 trials) for both WEAT test sets used. We re-run the built-in TOP-1 analogy test on the final SC-GloVe model and find that it similarly achieves ~35%.

The conceptual novelty of this work lies in its inclusion of explicit bias representation, proposing the notion that an **ideal debiased model is not one which has no conception of bias, but rather one that is aware of its own bias and therefore can self-correct.** In the humanities, source criticism is a method of evaluating the contextual lens of an informational source in order to determine its reliability, and we believe this framework has relevance for the mitigating bias in NLP methods.

Future Work

One weakness of this approach is that it is highly specific to the chosen bias metric. The source criticism method we introduce must be performed for each of the WEAT word sets, which hinders a blanket applicability to remove all enumerated biases in a GloVe model at once.

Future work will seek to expand the coverage of the method to debias with respect to multiple metrics at once, in addition to incorporating statistical significance testing across a variety of different GloVe pre-training parameters and corpora.

Works Cited

[1] Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014
[2] Brunet, M. E., Alkalay-Houlihan, C., Anderson, A., and Zemel, R. Understanding the origins of bias in word embeddings.*36th International Conference on Machine Learning, ICML 2019*, 2019-June:1275–1294, 2019.
[3] Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, apr 2017. ISSN 10959203. doi: 10.1126/science.aal4230.
[4] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
[5] Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. *34th International Conference on Machine Learning, ICML 2017*, 4:2976–2987,2017.
[6] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Account-ability, and Transparency*, pp. 648–657, 2020.

Acknowledgements
The author would like to thank Dr. Marcus Tomalin who provided feedback on an earlier version of this work, and Umang Bhatt, who provided numerous helpful discussions.