A Source-Criticism Debiasing Method for GloVe Embeddings

Hope McGovern
Newnham College
hem52@cam.ac.uk

Abstract

It is well-documented that word embeddings trained on large public corpora consistently exhibit known human social biases. Although many methods for debiasing exist, they almost all fixate on completely eliminating biased information from the embeddings, and often diminish training set size in the process. In this paper, we present a simple yet effective method for debiasing GloVe word embeddings (Pennington et al., 2014) which works by incorporating explicit information about training set bias rather than removing biased data outright. Our method runs quickly and efficiently with the help of a fast bias gradient approximation method from Brunet et al. (2019). As our approach is akin to the notion of 'source criticism' in the humanities, we term our method Source-Critical GloVe (SC-GloVe). We show that SC-GloVe reduces the effect size on Word Embedding Association Test (WEAT) sets without sacrificing training data or TOP-1 performance.

1 Introduction

Although many debiasing methods have been proposed to combat the problem of undesirable word associations in embeddings, each one suffers from their particular drawbacks and sometimes succeed only in hiding the latent word associations in latent space (Gonen and Goldberg, 2019). Debiasing remains a challenge because it is not well understood how individual training data or subset of datasets influence models down the line. However, recent work in applying influence functions to word embeddings has made it computationally tractable to identify how much each individual training example influences the overall bias of the model at inference time. We make use of this development to re-embed GloVe word vectors by artificially scaling the co-occurrence matrix so the model learns stronger word relationships from documents that are not biased (with respect to some predefined bias metric).

Our approach is based on the notion that an ideal debiased model is not one that has no conception of bias, but rather one that understands its own bias and therefore can self-correct. This is essentially an inclusion of an explicit bias awareness. 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 apply that concept here.

Computationally, we use a method of approximating differential bias from Brunet et al. (2019) to generate a weighting factor for each document which corresponds to how much it affects downstream bias at test time. This is accomplished with a single pass through the corpus. We then use these weighting factors to update the word-vectors that are relevant to our bias metric to what they would have been if the biased document had been counted as 'less reliable' during training. This is essentially the inclusion of an explicit bias representation.

Our method is simple, elegant, and represents a novel approach of debiasing via explicit bias inclusion.

2 Background

2.1 Bias in Word Embeddings

Word embeddings are used widely in natural language processing (NLP) for a broad range of downstream tasks. However, as Caliskan et al. (2017) and Garg et al. (2018) have shown, these embeddings confirm many human social biases, some of which lead to problematic behaviors of machine learning models which rely on pretrained embeddings (Kiritchenko and Mohammad). Among popular embedding models are Word2Vec (Mikolov et al.), GloVe (Pennington et al., 2014), and FastText (Joulin et al., 2017). All three use

unsupervised learning techniques on billions of words sourced from large internet corpora such as Wikipedia, Common Crawl (CC) or the Google News corpus. Bolukbasi et al. (2016) demonstrated that a simple analogy task could be used to reveal unwanted latent semantic associations in the embeddings; for example, the response from pretrained GloVe embeddings to the question "man is to computer programmer as woman is to X" rendered "homemaker" as the most likely answer.

Caliskan et al. (2017) introduced a standardized method of measuring biases present in word embeddings from a template of the Implicit Association Test, which is used widely in the field of sociology to measure latent human prejudices; this includes everything from morally neutral biases, such as 'flowers are more pleasant than insects' to more problematic ones, such as 'European names are more pleasant than African names'. From this, they derive 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.

Some previously proposed debiasing methods happen post-training, such as Bolukbasi et al. (2016), focus on zeroing the 'gender direction', i.e. the projection of each word on a predefined gender direction written as $\mathbf{w} \cdot (\mathbf{he} - \mathbf{she}) = 0$. Other methods attempt to mitigate bias during training, such as Zhao et al. (2020)'s method of encouraging gendered information to reside in the tail end of the word vector during training and then removing the biased subset of the word vector before use in an NLP system.

2.2 Influence Functions

With the use of influence functions, a methodology borrowed from robust statistics, it is possible to determine which inputs to a model exert the most influence over model inference at test time (Koh and Liang, 2017). Fundamentally, influence functions 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. Recent work in this area has introduced computationally efficient tools for performing fast influence function calculations for large models (Guo et al., 2020) and even through a multi-stage training (pre-training and fine-tuning) process (Chen et al., 2020). These tools open up the possibility of better model understanding by being able to directly track how input data lead to downstream inference.

Brunet et al. (2019) apply an influence function based approximation method to word embeddings as a way of explaining learned bias. With their method, it is possible to determine the subsets or individual documents that are most responsible for disparate gendered representations in learned word vectors. They demonstrate this method by creating perturbation sets which remove a number of biasing documents and compare the predicted differential bias sets with ground truth removal and re-embedding. They show remarkable agreement between the two, giving strong evidence that their approximation function is a trustworthy proxy for differential bias.

3 Methodology

3.1 Choice of Bias Metric

For our experiments, we consider the *effect size* of two different WEAT bias word sets as introduced by Caliskan et al. (2017).The WEAT test itself measures the similarity of words a and b in word embedding w as measured by the cosine similarity of their vectors, $\cos w_a, w_b$. In WEAT1, the target word sets relate to *science* and *arts* terms, with the expectation that scientific terms will cluster more with 'male' attributes sets and art terms will cluster more with 'female' attribute sets. In WEAT2, the target word sets relate to *weapons* and *instruments*, with the expectation that weaponry terms will cluster more with 'unpleasant' attribute sets and musical terms will cluster more with 'pleasant' attribute sets. A full list of the word sets used is available in Table 1. Of the bias sets used, WEAT1 is broadly considered to reflect a problematic bias, while WEAT2 reflects one that is more benign.

3.2 GloVe

We focus on only one kind of embedding model, GloVe, in this paper, although our method could theoretically be applied to others. GloVe (Global Vectors for word representations) is essentially a log-bilinear model with a weighted least-squares objective (Pennington et al., 2014). The underlying intuition for the model is that words which occur in the same context more frequently within a given corpus are more likely to be semantically linked. The training objective of GloVe is to learn word

WEAT	Target Sets	Attribute Sets
I	**science:** science, technology, physics, chemistry, einstein, nasa, experiment, astronomy	**male:** male, man, boy, brother, he, him, his, son
	arts: poetry, art, shakespeare, dance, literature, novel, symphony, drama	**female:** female, woman, girl, sister, she, her, hers, daughter
II	**instruments:** @bagpipe, cello, guitar, lute, trombone, banjo, clarinet, harmonica, mandolin, trumpet, bassoon, drum, harp, oboe, tuba, bell, fiddle, harpsichord, piano, viola, bongo, flute, horn, saxophone, violin	**pleasant:** caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation
	weapons: arrow, club, gun, missile, spear, axe, dagger, harpoon, pistol, sword, blade, dynamite, hatchet, rifle, tank, bomb, firearm, knife, shotgun, teargas, cannon, grenade, mace, slingshot, whip	**unpleasant:** abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

Table 1: Full WEAT word lists

vectors such that their dot product equals the logarithm of the words' probability of co-occurrence.

Formally, this happens in two steps[1]. In the first, a sparse co-occurrence matrix $X \in \mathbb{R}^{V \times V}$ is extracted from the corpus. Each entry in the matrix, X_{ij} represents a weighted count of the number of times that word j occurs in the context window of word i. For the second step, the optimal embedding parameters w^*, u^*, b^*, and c^* are learned via gradient-based optimization such that they minimize the loss:

$$J(X, w, u, b, c) = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})^2 \tag{1}$$

where $w_i \in \mathbb{R}^D$ is the embedding of the ith word in the vocabulary. We use an embedding dimension of 75. The set of $u_j \in \mathbb{R}^D$ are the context word vectors, and b_i and c_j are the bias terms for w_i and u_j, respectively. $f(x)$, the weighting function, is used to attribute more importance to common word occurrences.

3.3 Data Description

Our dataset consists of a corpus constructed from a Simple English Wikipedia dump[2], using 75-dimensional word embedding vectors. The TOP-1 analogies test (shipped with the GloVe code base) measured approximately 35%, which is certainly lower than state-of-the-art but high enough for our purposes of demonstrating proof-of-concept of our method. We note that future work would ideally consider a broader range of corpora for testing. We train an initial GloVe embedding model on this corpus; the relevant corpus statistics and GloVe training hyperparameters are listed in Table 2.

Table 2: Experimental setup for Wiki corpus.

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%

3.4 Differential Bias

For our debiasing method, it is crucial that we know how much each document in the corpus contributes to its downstream effect size. The naive way to compute this would be to remove a single document from the corpus and completely retrain the embedding, but this is impractical and computationally intractable even for relatively small NLP corpora. However, Brunet et al. (2019) introduce a method for calculating an approximation of the resulting embedding change by applying a modified

[1] In the following sections, we use the notion provided in Brunet et al. (2019)

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

version of influence functions[3]. They approximate that the optimal word vector learned from the initial training, w_i^*, will change with respect to a given corpus perturbation (removing a document) as:

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

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 $\tilde{X}$ is the co-occurrence matrix discounting the co-occurrence matrix of document i. H is the Hessian with respect to only word vector w_i of the point-wise loss at X_i, and $\nabla_{w_i} L(X_i, w)$ is the gradient of the point-wise loss function at X_i with respect to only word vector w_i.

We use this method to get an approximation of the differential bias of each document in the Wiki corpus. Once that is calculated, we pass the co-occurrence matrix, the trained GloVe model, the WEAT test words, and the newly created list of differential biases, where $\beta^{(k)}$ is the differential bias approximation for document k to our method. The full algorithm may be seen in Algorithm 1. It bears similarity to the differential bias approximation in Brunet et al. (2019) with the crucial changes that we weight the subtractor of the co-occurrence matrix by the pre-calculated differential bias, and that we actually set the approximated vector as the vector stored in the final GloVe model. This is because we are essentially calculating what the word vector would have been had it paid less attention to a biased document, so we want to update the word vector to reflect that change, rather than just temporarily store it for comparison.

The intuition for this algorithm is that WEAT words which appear in heavily biased documents will lend the model unsavory associations between those words, so we want to maximize the strength of the connection between target and attribute words that appear together in unbiased or, better yet, de-biasing documents (those with a negative differential bias value).

Given a set of WEAT words in a document, there are three possible actions according to our method: 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. If the document increases bias down the line, the co-occurrence matrix is slightly artificially decreased for the relevant words, scaled by how heavily biased it is. If the document decreases bias downstream, the co-occurrence matrix is slightly increased for the relevant words, artificially strengthening the co-occurrences of more balances terms. We find this method to be an elegant and intuitive approach to debiasing word embeddings.

[3]the code for this is publicly available at https://github.com/mebrunet/understanding-bias

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 $\cap(S \cup T \cup A)$ **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 height

Table 3: WEAT Effect Sizes.

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

4 Results

We report our results in Table 3, which shows the baseline WEAT effect sizes compared to our model, SC-GloVe, where the best result is in boldface. We show that SC-GloVe does decrease the effect size for both WEAT test sets used. Each of these scores was averaged over 10 trials to counteract the variability that is inherent in the optimization process. We re-run the built-in TOP-1 analogy test on the final SC-GloVe model and find that it similarly achieves around 35%.

5 Discussion

Our intention in this paper is to present a compelling proof-of-concept for a novel method rather than to fine-tune it and benchmark a competitive performance against existing debiasing methods. As such as, we simply note that against our two chosen WEAT test sets, our method successfully decreases downstream bias effect size. This is done without removing training data, but by a simple

weighting factor and efficient re-embedding algorithm. Encouragingly, it does not decrease TOP-1 analogy performance, which indicates the the method is targets only the offending vectors.

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.

Due to time constraints, the statistical significance testing of our results is omitted. However, we do note that future work would include more rigorous testing of this method with a variety of different GloVe pre-training parameters and corpora. We are satisfied to present proof-of-concept on a relatively small corpus size in this paper.

6 Conclusion

We have presented a simple, effective, and conceptually novel method for incorporating explicit bias representation into a word embedding debiasing method. We demonstrate that our model, SC-GloVe, can reduce downstream bias effect sizes with respect to a defined bias metric. We hope to refine this method further with future work in order to demonstrate its value in debiasing word embeddings for NLP models.

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. 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.

Marc Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:1275–1294.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Hongge Chen, Si Si, Yang Li, Ciprian Chelba, Sanjiv Kumar, Duane Boning, and Cho Jui Hsieh. 2020. Multi-Stage Influence Function. *arXiv*, pages 1–16.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):E3635–E3644.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:609–614.

Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. 2020. FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. (2016).

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 67, pages 427–431, Stroudsburg, PA, USA. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. Technical report.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *34th International Conference on Machine Learning, ICML 2017*, 4:2976–2987.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. Technical report.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai Wei Chang. 2020. Learning gender-neutral word embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4847–4853.