An Interactive Visual Demo of Bias Mitigation Techniques for Word Representations From a Geometric Perspective*

Archit Rathore ARCHIT.RATHORE@UTAH.EDU
University of Utah, USA

Sunipa Dev SUNIPA@CS.UCLA.EDU
University of California, Los Angeles, USA

Vivek Srikumar SVIVEK@CS.UTAH.EDU
Jeff M Phillips JEFFP@CS.UTAH.EDU
University of Utah, USA

Yan Zheng YAZHENG@VISA.COM
Michael Yeh MIYEH@VISA.COM
Junpeng Wang JUNPENWA@VISA.COM
Wei Zhang WZHAN@VISA.COM
VISA Research, USA

Bei Wang BEIWANG@SCI.UTAH.EDU
University of Utah, USA

Editors: Douwe Kiela, Marco Ciccone, Barbara Caputo

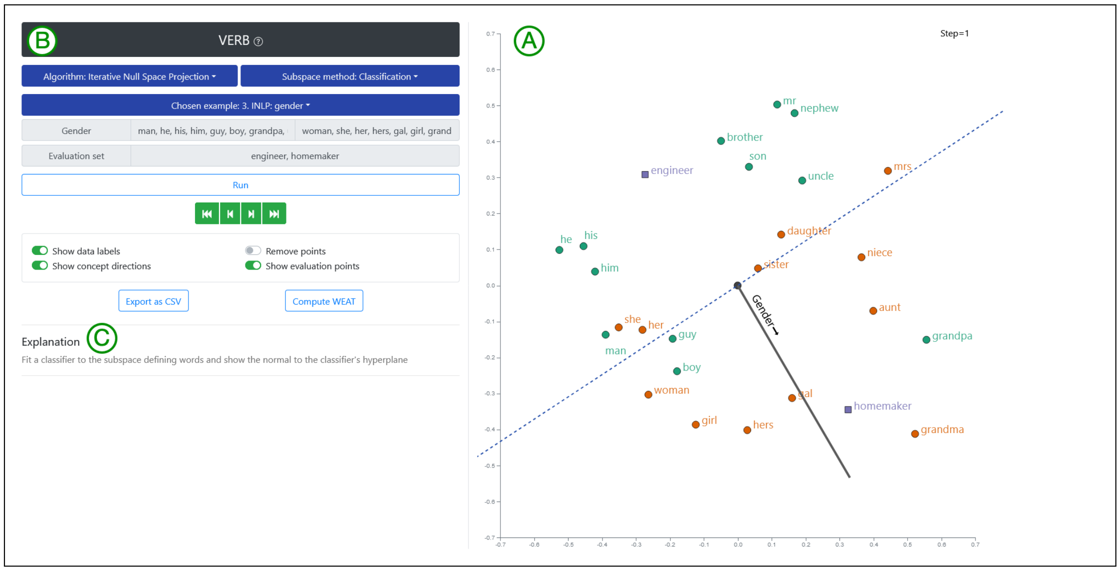
Abstract

Language representations are known to encode and propagate biases, i.e., stereotypical associations between words or groups of words that may cause representational harm. In this demo, we utilize interactive visualization to increase the interpretability of a number of state-of-the-art techniques that are designed to identify, mitigate, and attenuate these biases in word representations, in particular, from a geometric perspective. We provide an open source web-based visualization tool and offer hands-on experience in exploring the effects of these debiasing techniques on the geometry of high-dimensional word vectors. To help understand how various debiasing techniques change the underlying geometry, we decompose each technique into modular and interpretable sequences of primitive operations, and study their effect on the word vectors using dimensionality reduction and interactive visual exploration. This demo is primarily designed to aid natural language processing (NLP) practitioners and researchers working with fairness and ethics of machine learning systems. It can also be used to educate NLP novices in understanding the existence of and then mitigating biases in word embeddings.

Keywords: Visualization, debiasing, interactive data exploration and discovery, word representations, ethics

* This research was partially supported by NSF IIS-1816149, CCF-2115677, and Visa Research.

© 2022 A. Rathore, S. Dev, V. Srikumar, J.M. Phillips, Y. Zheng, M. Yeh, J. Wang, W. Zhang & B. Wang.



VERB

Algorithm: Iterative Null Space Projection

Subspace method: Classification

Chosen example: 3. INLP: gender

Gender	man, he, his, him, guy, boy, grandpa,	woman, she, her, hers, gal, girl, grand
Evaluation set	engineer, homemaker	

Run

● Show data labels

● Show concept directions

○ Remove points

● Show evaluation points

Explanation

Fit a classifier to the subspace defining words and show the normal to the classifier's hyperplane

Figure 1: With VERB, users can explore the high-dimensional word representations interactively before, during, and after applying bias mitigation techniques. (A) **Embedding View** highlights a subset of word embeddings using dimensionality reduction and visualizes the step-by-step transformations of their embeddings. (B) **Control Panel** enables users to configure each debiasing technique and provides controls to iterate through each step of the transformation. (C) **Explanation Panel** provides a description of the transformation.

1. Introduction

Distributed vector embeddings are becoming commonplace representations of elements from massive datasets. Perhaps the most common instance of this is the use of word vector embeddings like *Word2Vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014) across diverse natural language processing (NLP) tasks. Such representations are also commonly used to encode other discrete types of data such as graphs, spatial regions, and financial data. These vectorized representations directly capture similarities between these discrete objects, and also permit easy integration into machine learning tasks.

However, the high-dimensional and dense nature of these representations result in their overall lack of interpretability. As a result, it is hard to detect many complex biases which the representations are riddled with. In our context of word representations, *biases* refer to stereotypical associations between words or groups of words that may cause harm, e.g., the subordination of a demographic group. These biases are typically comprised of stereotypes about specific identities of persons (e.g., age, gender, race, etc.), and associate negative or polar connotations with them. For fair machine learning, it is often necessary to *modify* these vectorized representations to mitigate such biases (Bolukbasi et al., 2016; Dev and Phillips, 2019). These modifications use various approaches to identify and disentangle (Dev et al., 2021; Ravfogel et al., 2020) different concept subspaces from the embedding space. While effective, these methods can be hard to perceive, contrast, and understand due to the high-dimensionality of the representations.

In this demo, we present a tool—*VERB* (Visualization of Embedding Representations for deBiasing) (Rathore et al., 2021)—that allows a user to apply various debiasing methods in a step-by-step interactive manner. This enables the user to explore and understand which debiasing method is well-suited for their embeddings and analysis tasks. We focus on four debasing methods: *Hard Debiasing* (Bolukbasi et al., 2016), *Linear Projection* (Dev and Phillips, 2019), *Iterative Nullspace Projection* (INLP) (Ravfogel et al., 2020), and *Orthogonal Subspace Correction and Rectification* (OSCaR) (Dev et al., 2021). The VERB tool is open source, available at `https://github.com/tdavislab/verb`.

2. VERB Frontend and Backend

Frontend. The VERB frontend (Figure 1) allows real-time exploration of biases in word embeddings and provides a step-by-step illustration of how various debiasing algorithms attenuate those biases. The tool allows users to input any words from the embedding vocabulary. For the expediency of the demo, it also offers a variety of built-in examples to help a user get started, and to illustrate the core functionality. The interface is implemented using a web stack and can be accessed both locally and over the web.

Backend. In developing VERB, we decompose each debiasing technique into a three-step process. The first identifies a linear concept subspace among the vectorized representations that capture the direction of bias (e.g., of gender or nationality). The second uses this subspace to transform the representations in a simple and controlled way with the goal of attenuating the bias. The last evaluates the transformed representations.

Step 1: Subspace Identification: VERB allows for four commonly used methods of subspace identification, each resulting in a single unit vector. For each method, it requires the user to input seed words for identifying this vector, or use a pre-filled option that is available in the tool. *PCA* returns the top principal component of all words associated with the concept (Bolukbasi et al., 2016). *Paired-PCA* (Rathore et al., 2021) requires a set of pairs of words, and takes the top principal component of all difference vectors between all pairs. *Two-means* (Dev and Phillips, 2019) requires two sets of words, takes the mean of each set, and returns the unit vector from one mean to the other. *Classification* (Ravfogel et al., 2020) requires two sets of words, finds the best linear classifier of that set, and returns the normal direction to that separator.

Step 2: Bias Mitigation: VERB allows for four methods of removing bias associated with the identified subspace. *Linear Projection* (Dev and Phillips, 2019) projects all words away from the identified subspace. *Hard Debiasing* (Bolukbasi et al., 2016) starts with linear projection, but only applied to a subset of words associated with the concept direction. It then equalizes a set of user-provided word-pairs associated with the subspace. *Iterated Null Space Identification* (Ravfogel et al., 2020) applies linear projection, then re-identifies a potentially residual concept subspace, and repeats linear projection, etc, until no subspace can be identified. *OSCaR* (Dev et al., 2021) identifies two subspaces, and then performs a graded rotation that smoothly rotates all words so the two subspaces are orthogonal.

Step 3: Bias Evaluation: VERB applies the standard WEAT score (Caliskan et al., 2017) before and after the bias mitigation. Additionally, it allows a user to export the embedding after bias has been mitigated for a wide variety of evaluations or applications.

Implementation. VERB frontend is developed with HTML/CSS/Javascript stack and D3.js. Its backend is developed using Python and Flask. VERB is (by default) equipped with a 50-dimensional *GloVe* embeddings of the 100K most frequent words taken from the Wikipedia 2014 + Gigaword corpus (Pennington et al., 2014). It also provides a larger downloadable *GLoVe* embedding (300-dimensional with the 100K most frequent words) from the Common Crawl corpus (`https://commoncrawl.org/`).

3. Audience Interaction with the Demo at NeurIPS

The first phase of the audience interaction consisted of an overview of VERB's intuitive interface (Figure 1). We explained various algorithmic and visualization components of the tool. Following this, the users had two options to interact with the tool. The tool could be auto-installed on their own computer using the command line, which took less than a minute. Alternatively, the users could also use a web interface supported by our server, which allows 20-30 users simultaneously.

With the visual interface, the users could get started quickly with more than a dozen pre-loaded examples that demonstrate various forms of biases in word vector embeddings. The VERB tool guides the user through algorithmic mechanisms to mitigate these biases. The tool includes standard examples on gender-occupation biases, and also non-standard ones based on age, ethnicity, nationality, and royalty. Moreover, users can modify the words used in the examples, the techniques to identify concept subspaces (e.g., gender or nationality directions), and techniques to mitigate the associated biases. The visual interface allows users to easily compare different debiasing techniques, understand their powers and limitations, and become more aware of the underlying (potentially unwanted) associations included in these ubiquitous representations. The tool is self-explanatory after a simple introduction, and contains pre-loaded examples with explanations. The pre-loaded examples also help illustrate various critiques about debiasing approaches.

Use Cases. The demo contains a self-guided tour that walks the user through an example usage scenario. An example uses case showing a relationship between male-female gender and strong-helpless associations, and their removal with linear projection is illustrated in Figure 2 using screen shots. VERB smoothly interpolates between these steps.

Take-away from the demo. Users gain a visual intuition of how debiasing techniques for word embeddings work, with a focus on their geometric properties. By interacting with the pre-loaded examples, they could visually gauge the differences and similarities between various debiasing techniques, both in the input requirements and the ways that the word representations are modified. This intuition is contextualized using popular examples

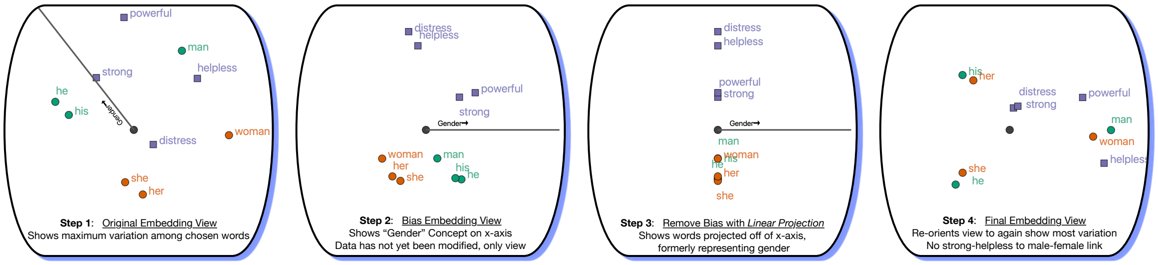

powerful
man
strong
helpless
he
his
Gender
distress
woman
she
her
Step 1: Original Embedding View
Shows maximum variation among chosen words
distress
helpless
powerful
strong
Gender
woman
her
she
man
his
he
Step 2: Bias Embedding View
Shows "Gender" Concept on x-axis
Data has not yet been modified, only view
distress
helpless
powerful
strong
Gender
man
his
woman
he
her
she
Step 3: Remove Bias with Linear Projection
Shows words projected off of x-axis,
formerly representing gender
his
her
distress
strong
powerful
man
woman
helpless
she
he
Step 4: Final Embedding View
Re-orients view to again show most variation
No strong-helpless to male-female link

Figure 2: VERB Screenshots of gender debiasing versus strong-distressed stereotype.

and a built-in evaluation metric. Finally, the users could generate their own exploratory experience by observing the stability of the methods by tweaking the input word sets, and by creating new use cases that highlight and mitigate biases in word embeddings or other high-dimensional representations.

References

T Bolukbasi, K W Chang, J Zou, V Saligrama, and A Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *ACM Transactions of Information Systems*, 2016.

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

Sunipa Dev and Jeff M. Phillips. Attenuating bias in word vectors. In *International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 879–887. PMLR, 2019.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Archit Rathore, Sunipa Dev, Jeff M. Phillips, Vivek Srikumar, Yan Zheng, Chin-Chia Michael Yeh, Junpeng Wang, Wei Zhang, and Bei Wang. VERB: Visualizing and interpreting bias mitigation techniques geometrically for word representations. *arXiv:2104.02797*, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.