Learned Gender Bias in Wikipedia Persists with Time

Blake Vente
Hunter College
Ralph.Vente09@myhunter.cuny.edu

Abstract

Word embeddings exhibit desirable properties when converting natural language to numerical vector representations. However, embeddings often internalize associations that parrot stereotypes pertaining to race, gender, and culture. Researchers have attempted to mitigate word embedding bias by altering the model after training, or changing the loss function, but to address bias comprehensively, Brunet et al. (2018) turn to the data where these biases originate.

1 Introduction

Distributional semantic models represent words with fixed-dimensional vectors based on the how words are used in context of other words in large quantities of unstructured text (Lison and Kutuzov, 2017). Word embedding models in particular, represent words as low-dimensional vectors intended to capture functional and topical relations between words (Lison and Kutuzov, 2017).

1.1 Learning Semantic Representations

One key advantage of word embedding models is that that they can be trained on large corpora of unstructured text. This gives the key advantage that retr Two prominent families of models that exhibit semantic representations of words are Continuous Bag-of-Words (CBOW) models and Skip-gram models.

In CBOW models, context is used to predict a target word. Formally, training example (x, y) is constructed by aggregating distributed representations of context words to form x, while the y. In particular, this "context" is defined as the k-word window about the target. If $k = 2$, and the aggregation function is a simple sum operation, then $x = \Sigma(w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2})$ and $y = w_i$.

1.2 Word Vector Arithmetic

If $T : V \to E$ denotes the embedding operation of word v from one-hot encoded (non-finite) vocabulary V into embedding space E with finite dimension $\dim E$.

Word embeddings have a wide variety of applications for natural language processing tasks including part-of-speech tagging, syntactic parsing, and named entity recognition (Lison and Kutuzov, 2017). In general, word embeddings are broadly applicable for any machine learning system that operates on vectors, including deep learning models and . Independently of their training mechanisms, word embeddings reproduce the bias intrinsic to the data they were trained on.

2 Word Embedding Bias

Reiterating $T : V \to E$ denotes the embedding operation of word v from vocabulary V into embedding space E with finite dimension $\dim E$, then we produce the well-documented and infamous expression $T(\text{computer programmer}) - T(\text{man}) + T(\text{woman}) \approx T(\text{homemaker})$ from (Bolukbasi et al., 2016b), which clearly exhibits the gender bias internalized by the learning algorithm that produced it. Hereafter, I denote the vector represntation of the word in upright boldface.

The presence of these dubious associations is not unique to one architecture or one configuration of hyper-parameters (Brunet et al., 2018). Thus, the inclusion of a word embedding step may compromise the integrity of downstream machine learning operations by injecting or accentuating this bias, making the process unsuitable for a wide range of applications. In sensitive domains such as granting loans, such model behavior may be illegal.

3 Prior Work in Debiasing

Many researchers have attempted to reduce bias in these embedding models, with limited success. Caliskan et al. (2017). These works can be partitioned ito two categories. First, there are those that alter the training process in some capacity. Then, there are those "post-processing" methods that alter vector representations at the end of training. Two such works follow: they define bias and attempt to minimize it without compromising performance on analogy tasks.

For example Bolukbasi et al. (2016a) formulate that the gender bias in a non-gendered word is its scalar projection on the $\vec{\textbf{he}} - \vec{\textbf{she}}$ axis. They zero out the first principal component in the gender direction. Gonen and Goldberg (2019) note that although the work was "extensive, thoughtful, and rigorous", the approach of Bolukbasi et al. (2016a) is inherently limited as the chosen definition of bias is hand-selected.

By contrast, Zhao et al. (2018) also attempt to mitigate historical biases in word embeddings, but they do so by altering the loss function of GloVe to concentrate onto the last element the component of the embedding most correlated with gender. Then, at inference time the last element of the embedding is truncated away an thus discarded, "encouraging" the representations of non-gendered words to be orthogonal to the gender direction. Gonen and Goldberg (2019) opine that this method carries the right intuition – the alterations the the model needs to happen at training time, but that the execution has limited efficacy. In fact, “indirect bias” is still very obvious even when “direct bias” is mitigated. They demonstrate that even simple models can still learn the latent biases in the word embedding. This suggests that the components of the vector that encode stereotypical notions of gender are distributed as a linear combination of many components, even if they are orthogonal to a primary “gender dimension”.

The general critique of both methods is that the definition of bias relating to distance to the gender direction Gonen and Goldberg (2019) believe that the correct method for removing gender bias, at a minimum, alters the training process.

Furthermore, Gonen and Goldberg (2019) remark that debiasing methods that don’t take the data into consideration merely "cover-up" the biases without addressing the full associations themselves.

4 WEAT: Operationalizing Bias

As the desire to de-bias models gained traction, there was little consensus on how to quantify bias, making it difficult to compare debiasing methods. Thus, Islam et al. (2016) developed the Word Embedding Association Test (WEAT) to quantify how word embeddings capture empirical information about the world from text corpora (Islam et al., 2016, 8). WEAT score was modeled after the Implicit Association Test (IAT) as a source of documented human biases (Islam et al., 2016, 2). Intuitively, WEAT is a generalization of the use of word embedding models to perform analogies. They first define $s(w, A, B) =$

$$\frac{\text{mean}_{a \in A} \cos(\hat{w}, \hat{a}) - \text{mean}_{b \in B} \cos(\hat{w}, \hat{b})}{\text{std}_{x \in A \cup B} \cos(\hat{w}, \hat{x})}$$

and then $S(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$. That is, S measures how close the associations between the target and attribute are. The more similar two sets of words are, the higher the WEAT score will be between them.

The minimum WEAT score is -2 and the maximum is 2.

4.1 Historical Biases in Word Embedding models are pervasive

The stereotypes reproduced by word embeddings are not just limited to gender, but also extend to race and culture.

This reinforces the claim that word associations that reproduce historical biases are learned the same way as analogies without bias.

5 Experimental Setup

To examine the behavior of historical biases in word embeddings over time, I train GloVe and FastText word embedings using the Gensim library by Řehůřek and Sojka (2010).

How does performance on analogies change with context window size increase?

I define Window Sizes $W = \{1, 5, 10, 15, 20, 25, 30\}$ and architectures $W = \{\text{FastText}, \text{Word2Vec}\}$

Claim: word embedding models learn bias the same way they do analogies.

transitivity

6 Implications

7 Limitations

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard S. Zemel. 2018. Understanding the origins of bias in word embeddings. *CoRR*, abs/1810.03611.

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

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. *CoRR*, abs/1903.03862.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

Pierre Lison and Andrey Kutuzov. 2017. Redefining context windows for word embedding models: An experimental study. *arXiv preprint arXiv:1704.05781*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings.