Reducing gender bias in word embeddings

CS229 Machine Learning, Stanford University

Introduction

Word embedding is a popular framework that represents text data as word vectors. These vectors capture semantics in language and are used in a variety of NLP and machine learning contexts. Recent research [1] has measured significant gender bias in popular word embeddings. This bias is believed to be amplified in applications that use the embedding. After measuring bias in embeddings produced by the popular GloVe [2] algorithm trained on an ordinary language corpus, we provide a methodology for modifying the algorithm to mitigate bias in the embedding while maintaining important relationships between gendered words that do not exhibit stereotype.

Measuring Bias

Starting with the 1 Billion Word Language Model Benchmark we applied an open source Cython implementation of the GloVe algorithm to produce a word embedding. To understand gender bias in word embeddings, we measure the direct and indirect bias by known metrics [1] as follows:

Direct Bias

We take a selected set of gendered pairs such as *she-he* from the word embedding and use PCA to identify a direction *g* that explains most of the variance among these pairs. Direct bias is measured as the average of the cosine similarity between words in a set of identified gender neutral words *N* and this direction *g*.

Indirect Bias

We also consider associations between gender neutral words. For example, if we project the *receptionist* along the *softball-football* direction, we see it is significantly more closely related to the word *softball* than *football* as an indirect consequence of gender stereotype. We measure indirect bias β between any gender neutral word pair *(w,v)* as follows

$$\beta(w,v) = \left(w \cdot v - \frac{w_\perp \cdot v_\perp}{\|w_\perp\| \cdot \|v_\perp\|}\right) / w \cdot v$$

$$w_\perp = w - (w \cdot g)g, \quad v_\perp = v - (v \cdot g)g$$

most "football" words	gender portion (%)	most "softball" words	gender portion (%)
midfielder	1	seamstress	18
captain	4	ballerina	10
footballer	0.01	podiatrist	42
mayor	12	salesperson	35
publisher	24	caregiver	26
architect	18	receptionist	62

Reducing Bias

Approach 1: Scaling the co-occurrence matrix entries

First, some notation. The co-occurrence matrix is denoted by X, whose entries X_{ij} count the number of times word j occurs in the context of word i. We use X_i to denote the sum over i, which is the number of times any word occurs in the context of i. Consider gender pairs such as (*she*, *he*), (*sister*, *brother*), (*woman*, *man*), (*Mary*, *John*). If an occupation word like *receptionist* is unbiased, it should co-occur across each pair with equal frequency. Disparity in co-occurrence counts (for example, if *boss* occurs 3x more frequently in the context of *he* than *she*) will contribute to biases in the word embedding. To correct for this, we shift the co-occurrence entries as follows (where *i* and *j* are gender pair words, and *k* is an occupation word):

$\frac{X_{ik}+s}{X_i} = \frac{X_{jk}-s}{X_j}$. Solving for s gives $s = \frac{X_i X_{jk} - X_j X_{ik}}{X_i + X_j}$ and $\beta_{ik} = \frac{X_{ik}+s}{X_i}$, $\beta_{jk} = \frac{X_{jk}-s}{X_j}$

The modified objective function for GloVe then is:

$$J = \sum_{i,j=1}^{V} f\left(\beta_{ij} X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(\beta_{ij} X_{ij})\right)^2$$

Approach 2: Regularization

The second approach modifies the objective function using regularization. We first compute word embeddings on our gender pairs, then use PCA to compute the gender direction *g*. Then we compute the rest of the word embeddings, this time adding a reugularization term which penalizes cosine similarity between our word vector and the gender direction. Here we define cosine similarity as follows:

$$\cos(u,v) = \frac{\vec{u} \cdot \vec{v}}{\|u\| \cdot \|v\|}$$

In this case, our modified objective function is:

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij})\right)^2 + \lambda \cos(w_i, g) + \gamma \cos(\tilde{w}_j, g)$$

For lambda and gamma we found there was a tradeoff between bias reduction and performance on the analogy test. We used a value of 10 for both parameters. For reference, the original GloVe objective function is below:

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij})\right)^2$$

Results

To quantitatively measure our results, we compare the direct bias [1] and MSR-analogy score [3] for the word embedding produced by the original GloVe algorithm and our modifications.

	Direct Bias	MSR-analogy
GloVe	0.11	0.37
Approach 1	0.055	0.34
Approach 2	**0.019**	**0.32**

Our second approach was the most successful with no major loss in the quality of the embedding. While we do not have a ground truth for indirect bias, we observe some desirable qualitative improvements. Focusing again on occupation words projected on the *softball-football* direction, the "most extremely softball" words are now *journeyman* and *game* which are relevant and do not exhibit gender bias. Stereotypical relationships between *softball* and words such as *receptionist*, *seamstress*, and *ballerina* have become more neutral on this axis.

Discussion

We have shown modifications can be made to the GloVe framework to make it more robust to gender bias. We believe this is a promising step to reducing gender discrimination that can result from the use of word embeddings produced today, which exhibit significant, measureable gender bias.

Given more time and resources we would further quantify our results by crowdsourcing analysis of gender stereotypes in top analogies to identify what percent of top analogies still exhibit perceived gender bias, similar to the approach taken in [1]. We also believe that interesting extensions of this research might focus on other sources of bias in word embeddings, such as racial bias, which have been targeted in other areas of machine learning, or by focusing on other popular word embedding frameworks like word2vec.

References

[1] T. Bolukbasi, K.W. Chang, J.Y. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.

[2] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[3] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013.

Gabrielle Badie (gab47@stanford.edu) • Tuhin Chakraborty (tuhin@stanford.edu) • Brett Rudder (brudder@stanford.edu)