Removing Bias from Word Embeddings

Fairness in AI: A replication study

Kylian van Geijtenbeek Thom Visser
Martine Toering Iulia Ionescu

MSc Artificial Intelligence
University of Amsterdam

January 31, 2020

Introduction

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). **Man is to computer programmer as woman is to homemaker? debiasing word embeddings.**

Claims from paper

- Gender bias in word embeddings
- Bias removed with debiasing algorithms (2)
- Performance of embeddings does not deteriorate
- Similar biases in other publicly available embeddings

Introduction

Popular word embeddings:

- Word2vec
- GloVe
- FastText

Introduction

Word2vec example: *dog*

- The man was walking his dog, when he tripped.
- The dog is chasing the cat.
- The dog barked a few times but then ate his food.

Context words: man, walking, tripped, chasing, cat, barked, ate, food

Introduction

Word2vec example: *dog*
Context words: man, walking, tripped, chasing, cat, barked, ate, food
Not context words: girl, car, tree, dishwasher, table, laptop, water, war

Positive samples:

- (dog, man)
- (dog, walking)
- (dog, tripped)
- (dog, cat)
- ...

Negative samples:

- (dog, girl)
- (dog, car)
- (dog, tree)
- (dog, dishwasher)
- ...

Introduction

Word2vec example: *dog*

Let each word be a d-dimensional vector, e.g.:

$$dog = [0.0134, 0.0692, 0.0273, \dots]$$
$$man = [0.0621, 0.0074, 0.0922, \dots]$$

We call this the *embedding* of a word.

Now we train the embeddings such that:

- the *cosine similarity* between two words in a *positive* sample is *high*.
- the *cosine similarity* between two words in a *negative* sample is *low*.

Method

Claim 1: There is gender bias in word embeddings

- Qualitative analysis:
  - Inspect analogies (male vs female)
  - Inspect projection professions on gender subspace
- Quantitative analysis: WEAT score [-2, 2]

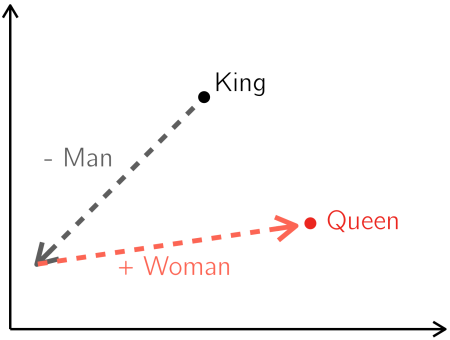

Figure 1: Vector differences between words represent relationship.

Method

Claim 2: This bias can be removed with their debiasing algorithms (2)

1. Identify gender subspace

- calculate means of defining sets (Equation 1)
- calculate SVD ($\boldsymbol{C}$) (Equation 2)
- k-dimensional gender subspace B is first k rows of SVD ($k = 1$ to match original paper)

$$\mu_i := \sum_{w \in D_i} \frac{\vec{w}}{|D_i|} \tag{1}$$

$$\boldsymbol{C} := \sum_{i=1}^{n} \sum_{w \in D_i} \frac{(\vec{w} - \mu_i)(\vec{w} - \mu_i)^T}{|D_i|} \tag{2}$$

Method

Claim 2: This bias can be removed with their debiasing algorithms (2)

2a. Hard debiasing (neutralize and equalize)

- gender neutral words shifted to zero in the gender subspace (neutralized) by subtracting projection of neutral word embedding vector onto gender subspace and renormalizing resulting embedding to unit length
- embedding is equalized, gender-pairs (*princess-prince*) adjusted in a way that all gender neutral words are equidistant to both female and male word in pair

$$\mu := \sum_{w \in E} \frac{\vec{w}}{|E|} \tag{3}$$

$$\vec{w} := \mu - \mu_B + \sqrt{1 - ||\mu - \mu_B||^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||} \tag{4}$$

Method

Claim 2: This bias can be removed with their debiasing algorithms (2)

2b. Soft debiasing

$$\min_T ||(TW)^T(TW) - W^TW||_F^2 + \lambda||(TN)^T(TB)||_F^2 \tag{5}$$

Optimization problem:

$$\min_X ||\Sigma U^T(X - I)U\Sigma||_F^2 + \lambda||N^TXB||_F^2 \qquad s.t. X \succeq 0. \tag{6}$$

T debiasing transformation
W matrix of all embedding vectors
λ hyperparameter to balance bias removal and conservation
N matrix of embedding vectors of gender neutral words
B gender subspace
$X = T^TT$
Σ diagonal matrix after SVD on W
$W = U\Sigma V^T$, U and V orthogonal matrices

Method

Claim 3: The performance of the embeddings does not deteriorate after using these algorithms

- Similarity benchmarks: RG-65 and WS-353
- Analogy benchmark: MSR

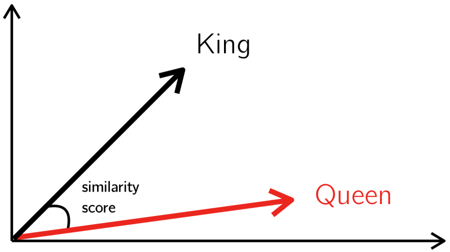

Figure 2: Similarity benchmarks.

Method

Claim 4: There are similar biases in other publicly available embeddings

- Word2vec
- FastText
- GloVe

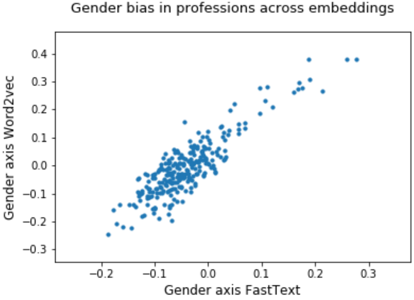

Figure 3: Similar occupational bias between word embeddings Word2vec and FastText. Datapoints represent occupation words.

Results and Discussion

Qualitative analysis bias:

FastText	
Professions closest to *she*	
nurse	librarian
socialite	dancer
housekeeper	singer
receptionist	vocalist
Professions closest to *he*	
inventor	commander
pundit	electrician
carpenter	footballer
headmaster	architect
Analogies woman : man	
supermodel:footballer	
beautiful:brilliant	

Table 1: Examples of bias present in analogies created from the word embeddings FastText before debiasing as well as gender bias in relation to occupations

Results and Discussion

Quantitative analysis bias and Performance of embeddings:

Word2vec	RG-65	WS-353	MSR	WEAT
Before	77.7	68.8	46.8	1.5
Hard-debiased	77.5	68.5	47.0	0.4
Soft-debiased	77.7	68.8	46.8	-0.1

GloVe	RG-65	WS-353	MSR	WEAT
Before	83.1	66.4	37.5	1.7
Hard-debiased	83.4	66.6	37.6	0.5
Soft-debiased	83.1	66.4	37.4	0.8

FastText	RG-65	WS-353	MSR	WEAT
Before	83.9	74.1	55.9	1.5
Hard-debiased	83.5	74.2	56.0	0.5
Soft-debiased	84.3	74.1	54.7	0.4

Table 2: Performance of small sets of Word2vec, GloVe and FastText embeddings and WEAT before and after debiasing.

Conclusion

Claims from paper

- Gender bias in word embeddings
- Bias removed with debiasing algorithms (2)
- Performance of embeddings does not deteriorate
- Similar biases in other publicly available embeddings

Conclusion

Claims from paper

- Gender bias in word embeddings ✓
- Bias removed with debiasing algorithms (2)
- Performance of embeddings does not deteriorate
- Similar biases in other publicly available embeddings

Conclusion

Claims from paper

- Gender bias in word embeddings ✓
- Bias removed with debiasing algorithms (2)
- Performance of embeddings does not deteriorate
- Similar biases in other publicly available embeddings

Conclusion

Claims from paper

- Gender bias in word embeddings ✓
- Bias removed with debiasing algorithms (2)
- Performance of embeddings does not deteriorate ✓
- Similar biases in other publicly available embeddings

Conclusion

Claims from paper

- Gender bias in word embeddings ✓
- Bias removed with debiasing algorithms (2)
- Performance of embeddings does not deteriorate ✓
- Similar biases in other publicly available embeddings ✓