Adversarial Debiasing Word Embeddings

Rochelle Choenni (10999949)
Maximilian Filtenborg (11042729)
Victoria Foing (11773391)
Gaurav Kudva (12205583)

Overview

Topics today:

1. Adversarial Debiasing
2. Gender biased word embeddings
3. Method
4. Experiments
5. Results
6. Reproducibility
7. Discussion & Conclusion

Zhang et al. Mitigating unwanted biases with adversarial learning.

ICU Dataset Experiment

Biased word embeddings analogy Task:

1. **He : she :: king : ?**
2. **man : woman :: boss : ?**

Analogy 'formula':

$$v = x_2 + x_3 - x_1$$

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. ACM, 335–340.

Gender Biased Word Embeddings

man : woman :: boss : ?

biased	
neighbor	**similarity**
bosses	0.577
girlfriend	0.529
waitress	0.509
woman	0.506
abruze	0.498
gaigalas	0.498
prostitute	0.497
backtalking	0.492
gambarro	0.49

Adversarial Debiasing

A debiasing framework for any machine learning model.

Adversarial debiasing is a way of debiasing using a Adversary model, similar to a GAN setup.

Premise: There is a protected variable z, that should not be incorporated in the prediction

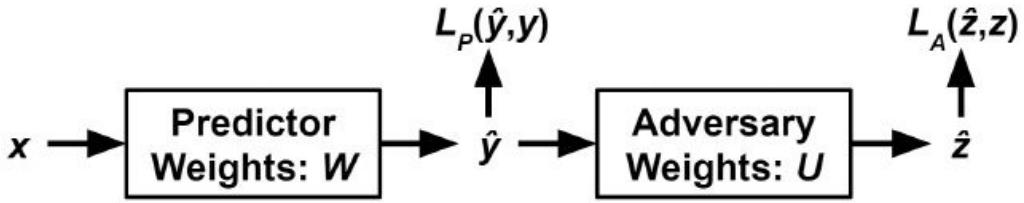


Figure 1: The architecture of the adversarial network.

Predictor & Adversary models

Predictor predicts the analogy word embedding.

Adversary predicts the **protected** variable.

Question: How do we quantify **z**?

Algorithm 2 predictor_model(self, features):

$d = self.embedding\_dim$
$x1 = features[:, 0 : d]$ // split features into 3 word vectors
$x2 = features[:, d : d \cdot 2]$
$x3 = features[:, d \cdot 2 : d \cdot 3]$
$v = x2 + x3 - x1$ // analogy vector
$\hat{y} = v - self.w \cdot self.w^T v$ // PyTorch linear layer
return $\hat{y}$

Algorithm 3 adversary_model(self, $\hat{y}$):

$\hat{z} = u^T \hat{y}$ // PyTorch linear layer
return $\hat{z}$

Gender Subspace

We define a gender subspace.

The gender subspace is computed by the PCA of the word-pair differences, and by taking the first principal component with the largest eigenvalue.

10 male-female word pairs
["he", "she"], ["his", "her"], ["man", "woman"]
["himself", "herself"], ["son", "daughter"], ["father", "mother"]
["guy", "gal"], ["boy", "girl"], ["male", "female"], ["john", "mary"]

Table 1: 10 male-female word pairs

Algorithm 1 obtain_gender_subspace(pairs, k):

$diff$ = [x[0] - x[1] for x in pairs]
$pca = PCA(n\_components = k)$
$pca.fit(diff)$
return $pca.components\_$

Original results from paper

- After debiasing
  - Subservient roles are replaced by doctor professions
  - Correlation metric increases

biased		debiased	
neighbor	similarity	neighbor	similarity
nurse	1.0121	nurse	0.7056
nanny	0.9035	obstetrician	0.6861
fiancée	0.8700	pediatrician	0.6447
maid	0.8674	dentist	0.6367
fiancé	0.8617	surgeon	0.6303
mother	0.8612	physician	0.6254
fiance	0.8611	cardiologist	0.6088
dentist	0.8569	pharmacist	0.6081
woman	0.8564	hospital	0.5969

Table 1: Completions for `he : she :: doctor : ?`

- Biased $w^T \cdot g$ = 0.08
- Debiased $w^T \cdot g$ = 0.55

Experiments

- Test the **generalizability** of the method across different word embedding types:
  - Wikipedia2Vec, GoogleNews and GloVe.
- Perform **gridsearch** to optimize the hyperparameter values for each embedding type
- Monitor the **correlation** between the learned predictor weights and the gender direction
  - $w^T \cdot g$

Evaluation: Analogy examples

- Compare the **top 10 nearest neighbours** for both the biased and debiased models.
- Verify that **correct gender relations** are still preserved after debiasing

Sexism traps

1. He : she :: doctor : ?
2. Man : woman :: boss : ?
3. He : she :: pilot : ?
4. He : she :: rich : ?
5. He : she :: intelligent : ?
6. He : she :: director : ?

Gendered analogies

1. Man : woman :: he : she
2. Man : woman :: king : queen

Hyperparameter Tuning

Embedding	$\|\|w\|\|$	biased $w^T \cdot g$	debiased $w^T \cdot g$	λ	α
GoogleNews	1.31	-0.09	-0.38	2^{-6}	1.0
Wikipedia2Vec	1.13	-0.02	-0.46	2^{-6}	1.0
GloVe	1.12	-0.08	0.37	2^{-6}	1.0

Here we see that after debiasing, the correlation metric increases!

Results: Gender Biased Word Embeddings

man : woman :: boss : ?

biased		debiased	
neighbor	**similarity**	**neighbor**	**similarity**
bosses	0.577	bosses	0.556
girlfriend	0.529	capofamiglia	0.498
waitress	0.509	mobster	0.464
woman	0.506	abruze	0.461
abruze	0.498	ingarao	0.461
gaigalas	0.498	umetaro	0.461
prostitute	0.497	alimzhan	0.46
backtalking	0.492	ENTITY/...	0.459
gambarro	0.49	mobsters	0.456

Table 3: Wikipedia2Vec: Completions for man : woman :: boss : ?

Results: GoogleNews

he : doctor :: she :

Biased Neighbour	Biased Similarity	Debiased Neighbour	Debiased Similarity
nurse	0.659	gynecologist	0.632
gynecologist	0.647	nurse	0.624
nurse_practitioner	0.626	nurse_practitioner	0.602
midwife	0.6	pediatrician	0.585
pediatrician	0.592	midwife	0.572
dermatologist	0.558	doctors	0.556
ob_gyn	0.556	ob_gyn	0.552
pharmacist	0.556	physician	0.548
doctors	0.554	dermatologist	0.548

Results: Glove

he : doctor :: she :

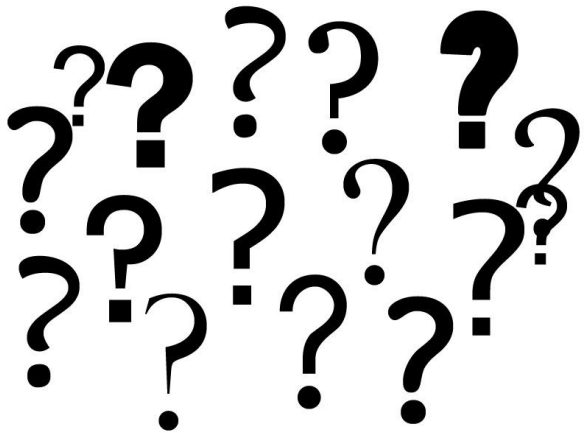
Biased Neighbour	Biased Similarity	Debiased Neighbour	Debiased Similarity
nurse	0.653	mordrid	0.544
doctors	0.652	gynecologist	0.485
physician	0.627	midwife	0.48
pregnant	0.618	kerish	0.467
pregnancy	0.589	obstetrician	0.464
she	0.584	nurse	0.462
midwife	0.568	pediatrician	0.454
her	0.566	naturopath	0.451
patient	0.565	dermatologist	0.45

Results

- We do not reproduce the **exact results** in the paper
- We do reproduce **general debiasing trends** for all word embedding types
- The results are highly influenced by **hyperparameters**
- Still **hard to measure** to what extent the debiasing really worked
- The lack of a **quantitative metric** is the biggest weak point in the paper

Reproducibility challenges

- Equations in the paper
- Loss functions
- Word embeddings
- Hyperparameters
- Implementation and unofficial implementations



Discussion

- How can we know debiasing is successful?
  - Regarding the boss analogy, mobster clearly contains a negative connotation.
- Sexism is a complex phenomena shaped by human judgement, which makes it hard to quantitatively evaluate.
- It is also a risky to measure qualitatively, and would require a carefully selected group from different backgrounds