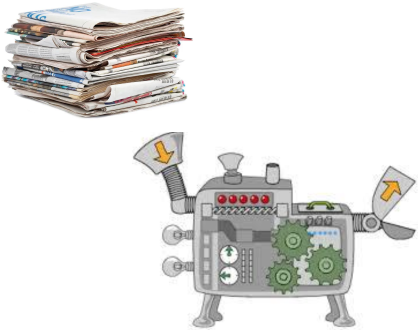
Interpretable Debiasing of Vectorized Language Representations with Iterative Orthogonalization

Prince Osei Aboagye[1], Yan Zheng[2], Jack Shunn[1], Chin-Chia Michael Yeh[2], Junpeng Wang[2], Zhongfang Zhuang[2], Huiyuan Chen[2], Liang Wang[2], Wei Zhang[2], Jeff M. Phillips[1]

[1]University of Utah [2]Visa Research

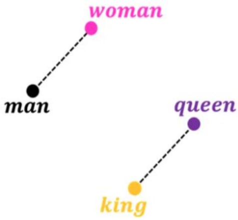
ICLR, May 2023

Word Embeddings



man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Word — Word embedding



Bias in Representation

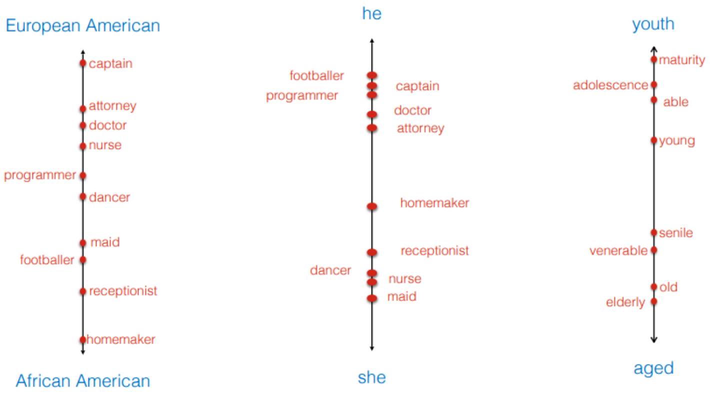
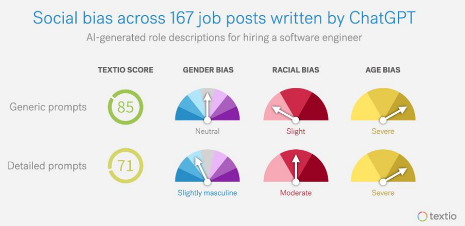


Image Credit: Dev, et al., 2021, "OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings"

Bias Amplification in ChatGPT



Source: https://textio.com/blog/chatgpt-writes-job-posts/99089591200

02-03-23 | WORKPLACE EVOLUTION

We asked ChatGPT to write performance reviews and they are wildly sexist (and racist)

Textio's cofounder Kieran Snyder observes that it takes so little for ChatGPT to start baking gendered assumptions into otherwise highly generic feedback.

Source: https://www.fastcompany.com/90844066/chatgpt-write-performance-reviews-sexist-and-racist

Daniel Munro
@dk_munro

ChatGPT: Historian of Philosophy.

"Name 10 philosophers"

1/6

Name 10 philosophers

1. Plato
2. Aristotle
3. Immanuel Kant
4. Friedrich Nietzsche
5. Jean-Jacques Rousseau
6. David Hume
7. René Descartes
8. Thomas Hobbes
9. John Stuart Mill
10. Søren Kierkegaard

2:01 PM · Mar 3, 2023 · **2.4M** Views

3,638 Retweets **860** Quotes **15K** Likes **2,016** Bookmarks

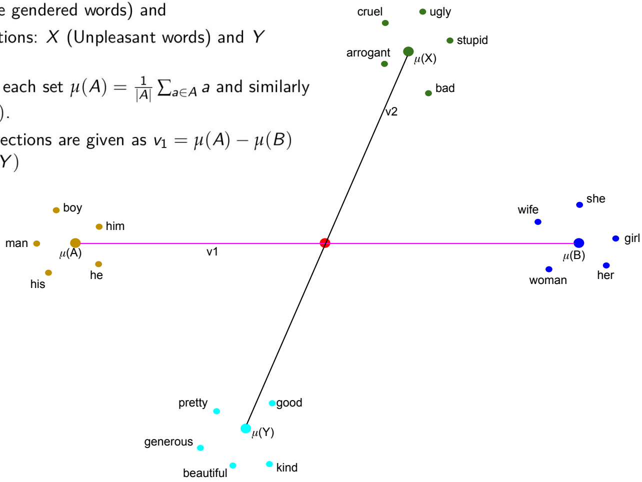
Source: https://mobile.twitter.com/dk_munro/status/1631761802500423680

Debiasing Representations

- Concept Subspaces Identification
- Debiasing and Disentangling of Subspaces

Concept Subspaces Identification: Two Means

- Given two pairs of concepts, Gender: A (male gendered words) and B (female gendered words) and
- Stereotypical associations: X (Unpleasant words) and Y (Pleasant words)
- We find the mean of each set $\mu(A) = \frac{1}{|A|}\sum_{a \in A} a$ and similarly for $\mu(B), \mu(X), \mu(Y)$.
- Then the concept directions are given as $v_1 = \mu(A) - \mu(B)$ and $v_2 = \mu(X) - \mu(Y)$



Debiasing and Disentanglement of Subspaces

- Linear Projection, LP (Dev & Phillips, 2019)
- Hard Debiasing, HD (Bolukbasi et al., 2016)
- Iterative Null Space Projection, INLP (Ravfogel et al., 2020)
- OSCaR (Dev et al., 2021)

Iterative Subspace Rectification

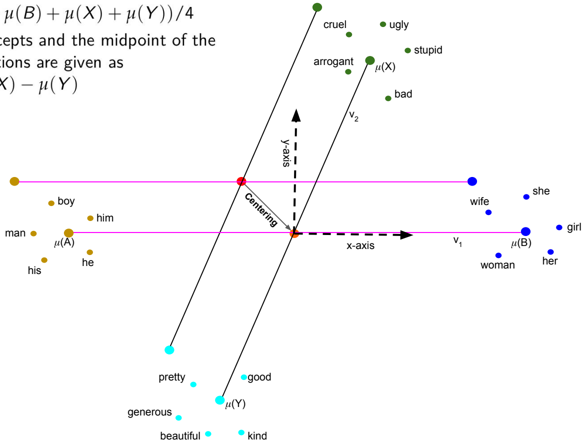
- In this work, we propose a new mechanism to augment a word vector embedding representation that offers:
  - ⋆ improved bias removal while retaining the key information
  - ⋆ resulting in the interpretability of the representation.
- We build on top of Orthogonal Subspace Correction and Rectification (OSCaR)
- We call our approach iterative subspace rectification (ISR), but add some subtle but significant modifications

Significant modifications to OSCaR

- Centering in ISR
- Rectification in ISR
- Uncentering in ISR
- Iteration in ISR

Centering in ISR

- Given $\mu(A)$, $\mu(B)$, $\mu(X)$ and $\mu(Y)$.
- We find the center $c = (\mu(A) + \mu(B) + \mu(X) + \mu(Y))/4$
- After centering each pair of concepts and the midpoint of the concept pairs, the concept directions are given as $v_1 = \mu(A) - \mu(B)$ and $v_2 = \mu(X) - \mu(Y)$



Rectification/Orthogonalization in ISR

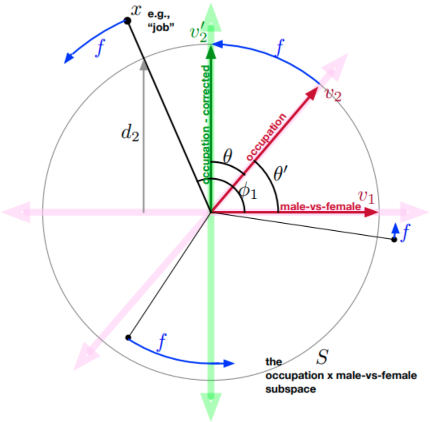


Image Credit: Dev, et al., 2021, "OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings"

Uncentering in ISR

- After rectification, we uncenter the orthogonal linear concept vectors.

Iteration in ISR

- We observe that the learned subspaces from OSCaR are not completely orthogonal
- As such, we iteratively run the entire centering, rectification, and uncentering process leading to our approach

Table 1: Dot Product Scores (dotP) on Gender Terms vs Pleasant/Unpleasant per iteration.

	Before	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5	Iter 6	Iter 7	Iter 8	Iter 9	Iter 10
dotP ISR	0.029	**0.007**	**0.002**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
dotP iOSCaR	0.029	0.128	0.204	0.340	0.532	0.716	0.535	0.731	0.473	0.686	0.667

Note: iOSCaR denotes iteratively running OSCaR

Evaluation of Debiasing and Rectification

- We evaluate the effectiveness of ISR in two ways:
  - ⋆ how well it actually orthogonalizes concepts
  - ⋆ how well it reduces bias

Evaluation using WEAT

Table 2: WEAT Score on Pairs of Concepts.

Concept1	Concept2	Orig.	LP	HD	INLP	OSCaR	SR	iOSCaR	ISR
Gen(M/F)	Career/Family	0.7507	0.7713	0.2271	0.3503	0.3343	0.3235	0.2154	**0.0114**
Gen(M/F)	Math/Art	0.7302	0.6975	0.1127	0.1262	0.5437	0.2928	0.4435	**0.0148**
Gen(M/F)	Sci/Art	1.1557	0.9068	0.1381	0.3776	0.8642	0.4245	0.5139	**0.0140**
Name(M/F)	Career/Family	1.7303	0.0421	0.0992	0.7916	0.8950	0.6556	0.3143	**0.0186**
Name(E/A)	Please/Un	1.3206	0.0800	**0.0518**	0.0960	0.3043	0.7015	0.0527	0.1678
Flower/Insect	Please/Un	1.3627	0.2395	0.1363	0.2713	0.6348	0.3957	0.1338	**0.0254**
Music/Weap	Please/Un	1.4531	**0.0373**	0.0942	0.0925	1.0135	0.4728	0.2043	0.0770

Evaluation using SEAT: Pre-trained Language Models

Table 3: SEAT test result (effect size) of gender debiased BERT and RoBERTa models. An effect size closer to 0 indicates less (biased) association.

Model	SEAT-6	SEAT-6b	SEAT-7	SEAT-7b	SEAT-8	SEAT-8b	Avg (↓)
BERT	0.931	0.090	−0.124	0.937	0.783	0.858	0.620
+ CDA	0.846	0.186	−0.278	1.342	0.831	0.849	0.722
+ DROPOUT	1.136	0.317	0.138	1.179	0.879	0.939	0.765
+ INLP	0.317	−0.354	−0.258	0.105	0.187	−0.004	0.204
+ SENTENCEDEBIAS	0.350	−0.298	−0.626	0.458	0.413	0.462	0.434
+ iOSCaR (Our approach)	0.931	0.078	−1.447	−1.178	−1.21	−1.491	1.056
+ ISR (Our approach)	0.048	−0.264	−0.253	−0.035	0.243	0.295	**0.190**
RoBERTa	0.922	0.208	0.979	1.460	0.810	1.261	0.940
+ CDA	0.976	0.013	0.848	1.288	0.994	1.160	0.880
+ DROPOUT	1.134	0.209	1.161	1.482	1.136	1.321	1.074
+ INLP	0.812	0.059	0.604	1.407	0.812	1.246	0.823
+ SENTENCEDEBIAS	0.755	0.068	0.869	1.372	0.774	1.239	0.846
+ iOSCaR (Our approach)	0.894	0.268	0.574	0.648	0.504	0.729	0.603
+ ISR (Our approach)	0.554	0.099	0.296	0.546	0.394	0.419	**0.385**

- We introduced a new mechanism for augmenting vectorized embedding representations, namely Iterative Subspace Rectification (ISR)
- Our approach:
  - ⋆ Offers improved bias removal while retaining the key concept information
  - ⋆ Can be extended to multiple concept subspaces
  - ⋆ Explicitly encodes concepts along the coordinate axis, making the resulting representations Interpretable

Thank you for your attention!