Interpretable Debiasing of Vectorized Language Representations with Iterative Orthogonalization

Prince Osei Aboagye

Visa Research

August 9, 2023

THE UNIVERSITY OF UTAH

VISA Research

High-dimensional Vectorized Embeddings

- Core element of the vast majority of machine learning tasks.
- Facilites learning, understanding concepts, and efficiently representing feature spaces.

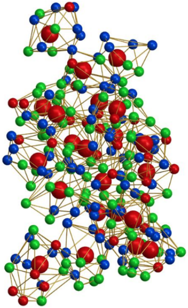


Image Credit: https://newsroom.ucla.edu/releases/first-ever-3d-atomic-imaging-amorphous-solid/

What are Embeddings?

- Mapped vector representations of data entities in high-dimension.

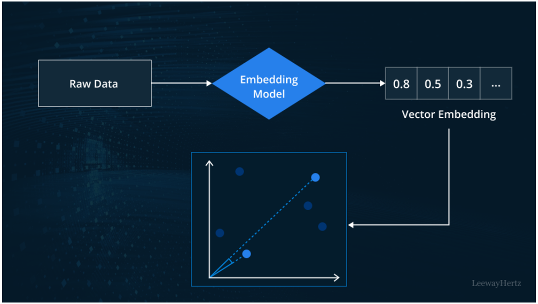


Image Credit: https://www.leewayhertz.com/embeddings-as-a-service/

Embedding Mechanism

- Self-supervised learning approach
- Effectively convey the meaning and structural relationships present in the input data.

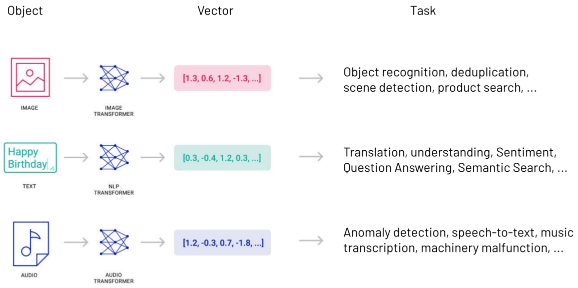
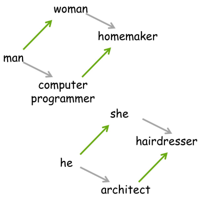


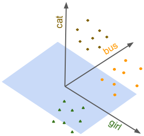
Image Credit: The Rise of Vector Data - https://www.pinecone.io/learn/rise-vector-data/

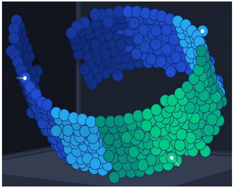
Challenges in High-dimension

- Difficult to think about or conceptualize the structure of embeddings in high-dimension.
- This makes analyzing and obtaining meaningful patterns within the embeddings difficult.

Inherent Challenges Associated with Embeddings







Bias

Lack of Interpretability

Structural Profiles

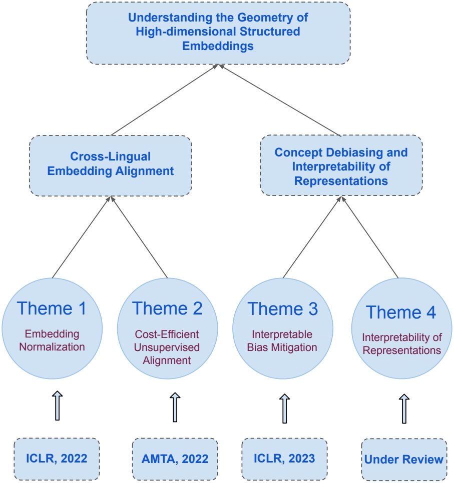
Way Forward: Geometric Approaches

- Despite these challenges, there are existing geometric techniques that can be used to gain insights and extract meaningful information from high-dimensional embeddings:
  - ⋆ Defining a basis
  - ⋆ Spectral structure through Eigen-Decomposition
  - ⋆ Normalization
- These techniques permit us to consider the geometry of these vectorized high-dimensional embeddings more appropriately.

Way Forward: Geometric Approaches

- Consequently, understanding the underlying geometrical structure of the high-dimensional vectorized embedding is of great interest.
- This motivates the central theme of my dissertation.

Central Theme



Central Theme: Publications

1. [ICLR, 2022] **P. O. Aboagye**, J. Phillips, Y. Zheng, J. Wang, C.-C. M. Yeh, W. Zhang, L. Wang, and H. Yang, *Normalization of language embeddings for cross-lingual alignment, in International Conference on Learning Representations*, 2022.
2. [AMTA, 2022] **P. O. Aboagye**, Y. Zheng, M. Yeh, J. Wang, Z. Zhuang, H. Chen, L. Wang, W. Zhang, and J. Phillips, *Quantized Wasserstein Procrustes alignment of word embedding spaces*, in Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas, 2022.
3. [ICLR, 2023] **P. O. Aboagye**, Y. Zheng, J. Shunn, C.-C. M. Yeh, J. Wang, Z. Zhuang, H. Chen, L. Wang, W. Zhang, and J. M. Phillips, *Interpretable debiasing of vectorized language representations with iterative orthogonalization*, in International Conference on Learning Representations (ICLR), 2023.
4. [**Under Review**] **P. O. Aboagye**, H. Pourmahmoodaghababa, Y. Zheng, C.-C. M. Yeh, J. Wang, H. Chen, L. W. Xin Dai, W. Zhang, and J. Phillips, *One-hot encoding strikes back: Fully orthogonal coordinate-aligned class representations*.

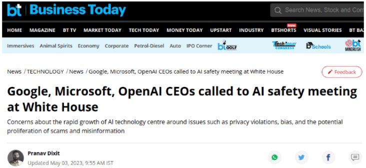
Theme 3

Interpretable
Bias Mitigation

⇑

[ICLR, 2023] **P. O. Aboagye,** Y. Zheng, J. Shunn, C.-C. M. Yeh, J. Wang, Z. Zhuang, H. Chen, L. Wang, W. Zhang, and J. M. Phillips, *Interpretable debiasing of vectorized language representations with iterative orthogonalization*, in International Conference on Learning Representations (ICLR), 2023.

AI Safety


bt Business Today

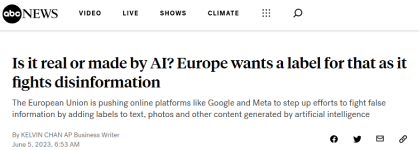
HOME MAGAZINE BT TV MARKET TODAY TECH TODAY MONEY TODAY UPSTART INDUSTRY BTSHORTS VISUAL STORIES

News / TECHNOLOGY / News / Google, Microsoft, OpenAI CEOs called to AI safety meeting at White House

Google, Microsoft, OpenAI CEOs called to AI safety meeting at White House

Concerns about the rapid growth of AI technology centre around issues such as privacy violations, bias, and the potential proliferation of scams and misinformation

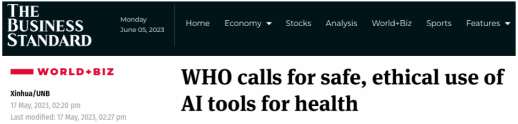
Pranav Dixit
Updated May 03, 2023, 9:55 AM IST


abc NEWS VIDEO LIVE SHOWS CLIMATE

Is it real or made by AI? Europe wants a label for that as it fights disinformation

The European Union is pushing online platforms like Google and Meta to step up efforts to fight false information by adding labels to text, photos and other content generated by artificial intelligence

By KELVIN CHAN AP Business Writer
June 5, 2023, 6:53 AM

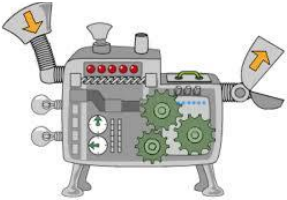

THE BUSINESS STANDARD Monday June 05, 2023 Home Economy Stocks Analysis World+Biz Sports Features

WORLD+BIZ

Xinhua/UNB
17 May, 2023, 02:20 pm
Last modified: 17 May, 2023, 02:27 pm

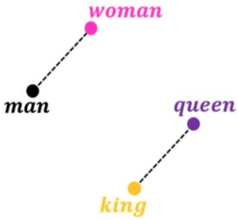
WHO calls for safe, ethical use of AI tools for health

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

$\underbrace{\qquad\text{}}_{\text{Word}}$ $\underbrace{\qquad\qquad\qquad\qquad\qquad\text{}}_{\text{Word embedding}}$



Bias in Language Representation

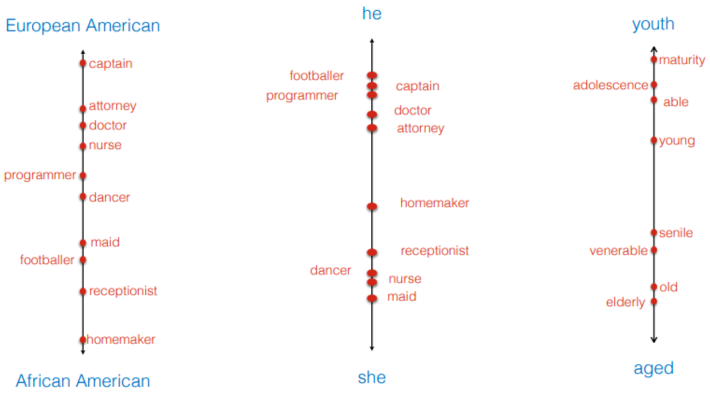
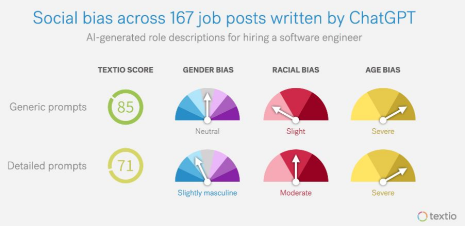


Image Credit: Dev, et al., 2021, "OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings"

Bias Amplification in ChatGPT



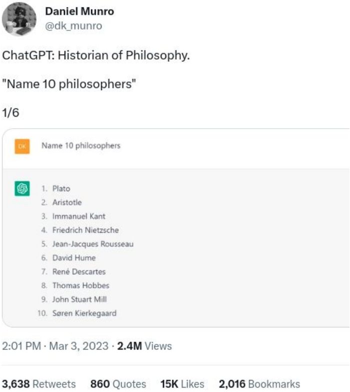
Source: https://textio.com/blog/chatgpt-writes-job-posts/99089591200


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

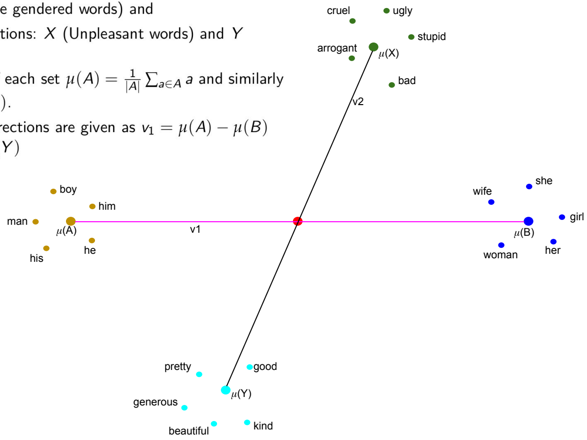
Source: https://mobile.twitter.com/dk_munro/status/1631761802500423680

Debiasing Representations by Post Processing

- 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 $v_2 = \mu(X) - \mu(Y)$

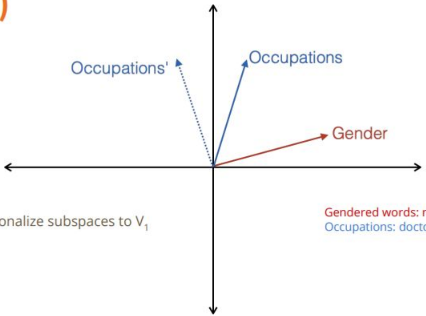


Debiasing and Disentanglement of Subspaces

- 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)

OSCaR

Orthogonal Subspace Correction and Rectification (OSCaR)



Orthogonalize subspaces to V_1 and V_2'.

Gendered words: man,woman, boy, he, lady, aunt...
Occupations: doctor, engineer, nurse, maid...

Our Proposed Method

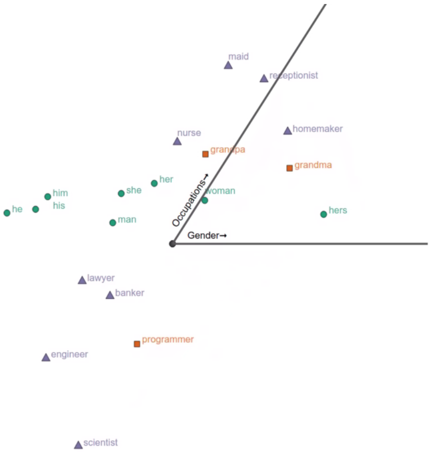
- In this work, we propose a new mechanism to augment a word vector embedding representation that offers:
  - ⋆ improved bias removal while retaining the concept information
  - ⋆ resulting in the interpretability of the representation.
- We build on top of Orthogonal Subspace Correction and Rectification (OSCaR)

Significant modifications to OSCaR

- Centering
- Rectification
- Uncentering
- Iteration

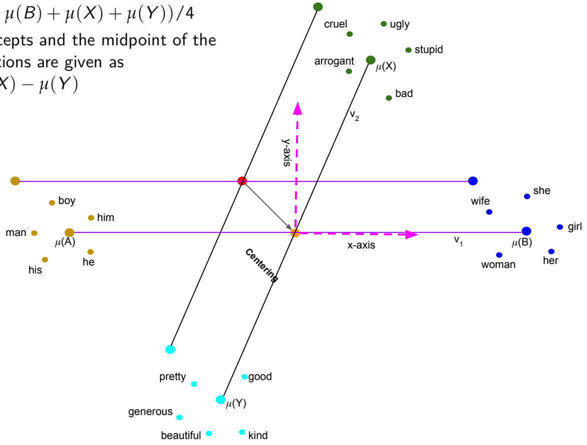
- We call our approach Iterative Subspace Rectification (ISR)

Point of Rotation in OSCaR

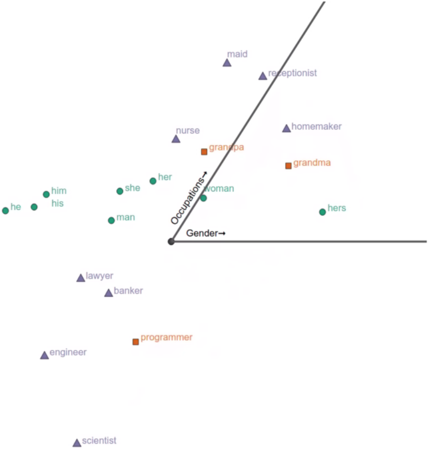


Centering in ISR

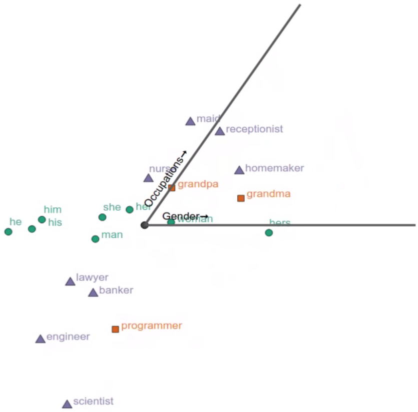
- Given $\mu(A)$, $\mu(B)$, $\mu(X)$ and $\mu(Y)$.
- We find the center $c = (\mu(A) + \mu(B) + \mu(X) + \mu(Y))/4$
- After centering each pair of concepts and the midpoint of the concept pairs, the concept directions are given as $v_1 = \mu(A) - \mu(B)$ and $v_2 = \mu(X) - \mu(Y)$



Example of Centering in ISR



Example of Centering in ISR



Rectification/Orthogonalization in ISR

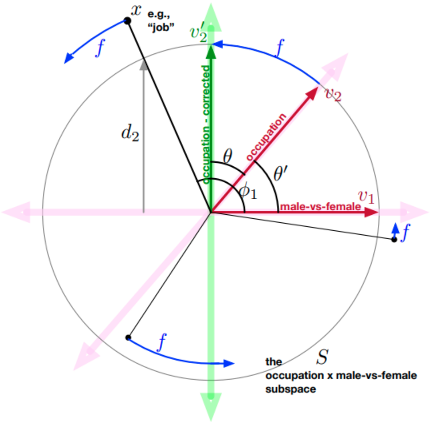
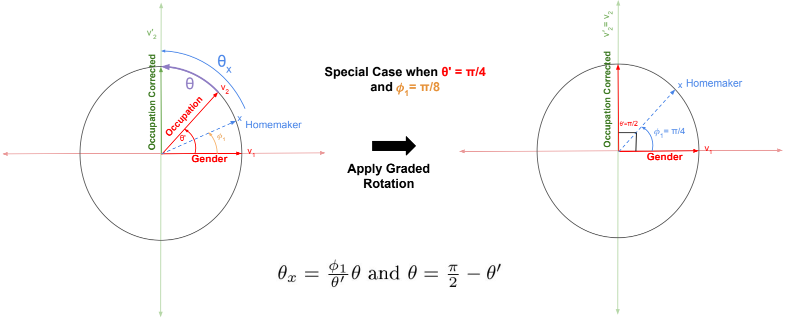


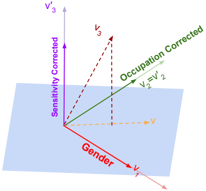
Image Credit: Dev, et al., 2021, "OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings"

Graded Rotation for Two Concept Subspaces

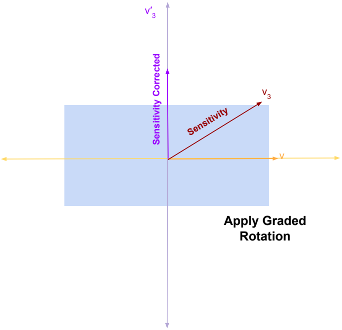


$$\theta_x = \frac{\phi_1}{\theta'}\theta \text{ and } \theta = \frac{\pi}{2} - \theta'$$

Graded Rotation for Three Concept Subspaces

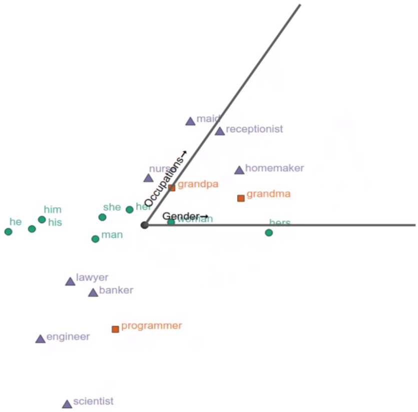


Now we consider the 2D subspace that is span by v and v_3

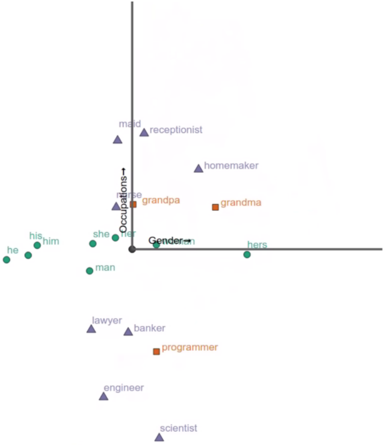


Apply Graded Rotation

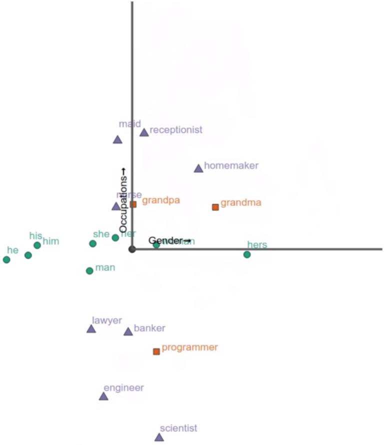
Example of Rectification in ISR



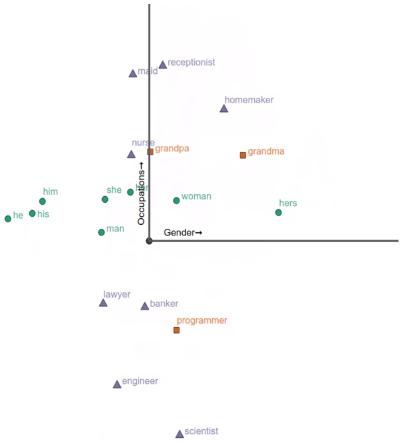
Example of Rectification in ISR



Example of Uncentering in ISR



Uncentering in ISR



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

Word Embedding Association Test (WEAT)

- $X = \{man, male, ...\}$ (definitionally male words
- $Y = \{woman, female, ...\}$ (definitionally female words)
- $A = \{programmer, engineer, scientist, ...\}$ (stereotypical male professions)
- $B = \{nurse, teacher, librarian, ...\}$ (stereotypical female professions)

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(a, w) - \frac{1}{|B|} \sum_{b \in B} \cos(b, w)$$

$$s(X, Y, A, B) = \frac{1}{|X|} \sum_{x \in X} s(x, A, B) - \frac{1}{|Y|} \sum_{y \in Y} s(y, A, B)$$

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

Self-WEAT (SWEAT) score

- $X = \{man, male, ...\}$ (definitionally male words
- $Y = \{woman, female, ...\}$ (definitionally female words)
- Randomly split X into X_1 and X_2
- Similarly split Y into Y_1 and Y_2
- Compute the WEAT score:

$$s(X_1, Y_1, X_2, Y_2)$$

Evaluation of Information Preserved

Table 3: SWEAT Score: Measuring Information Preserved.

Concept1	Concept2	Orig.	LP	HD	INLP	OSCaR	SR	iOSCaR	ISR
Gen(M/F)	Please/Un	1.7674	1.2685	1.1957	0.5528	1.5865	1.7678	0.6424	1.7683
Name(M/F)	Please/Un	1.9041	1.0893	1.9115	0.9475	1.8549	1.9046	1.2711	1.9044
Please/Un	Gen(M/F)	1.8762	0.0326	1.8862	0.7090	1.7810	1.8759	0.8006	1.8740
Career/Family	Gen(M/F)	1.8763	0.3530	1.8816	0.4549	1.7720	1.8733	0.7399	1.8527
Achieve/Anx	Gen(M/F)	1.8677	0.5435	1.8691	0.6893	1.7157	1.8694	0.3939	1.8705

Evaluation using SEAT

Table 4: 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**

3-concept Debiasing

Table 5: WEAT, dot product, and SWEAT scores for 3-concept debiasing among GT, NN, and P/U.

	WEAT			dot product			SWEAT		
Iteration	GT vs NN	GT vs P/U	NN vs P/U	GT vs NN	GT vs P/U	NN vs P/U	GT	NN	P/U
Orig.	0.1797	0.3337	1.1506	0.0589	0.0729	0.1721	1.7674	1.7289	1.8762
Iter 1	0.1157	0.1290	0.6195	0.0395	0.0273	0.0598	1.7692	1.7298	1.8768
Iter 2	0.0657	0.0442	0.3146	0.0252	0.0104	0.0204	1.7502	1.7459	1.8648
Iter 3	0.0316	0.0113	0.1974	0.0157	0.0041	0.0070	1.7637	1.7592	1.8715
Iter 4	0.0097	0.0015	0.1564	0.0096	0.0017	0.0024	1.7745	1.7711	1.8761
Iter 5	0.0040	0.0067	0.1423	0.0058	0.0007	0.0008	1.7545	1.7386	1.8603

Conclusion

- 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

Code

https://github.com/poaboagye/ISR

Out-of-Distribution Detection

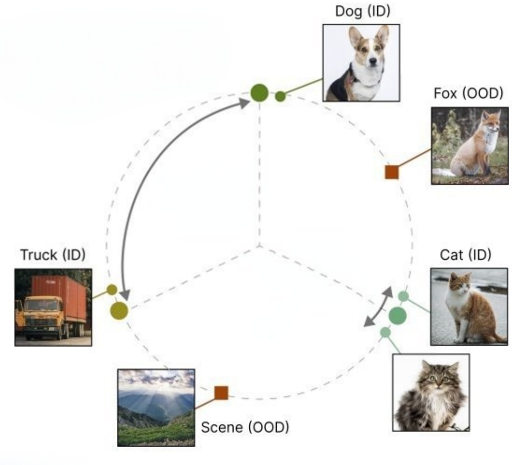


Image Credit: https://openreview.net/pdf?id=aEFaE0W5pAd

Convergence of Language and Vision Model Geometries

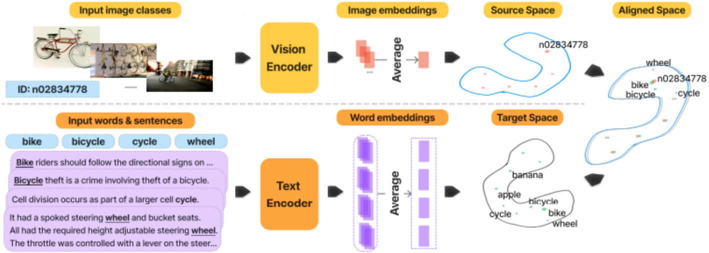


Image Credit: https://arxiv.org/pdf/2302.06555.pdf

Acknowledgement

Thank you for your attention!