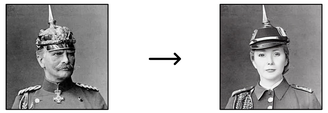
A Multidimensional Analysis of Social Biases in Vision Transformers

Jannik Brinkmann, Paul Swoboda, Christian Bartelt

International Conference on Computer Vision, 2023

Training data can mitigate, but not eliminate social biases

We measure the impact of augmenting the pre-training dataset with counterfactual images on reducing gender bias in both fine-tuning and pre-training phases of model training.



To generate counterfactual images:

1. Generate a caption (e.g., using CLIP).
2. Substitute terms in the caption using term pairs (e.g., “male” to “female”)
3. Use diffusion-based image editing with mask guidance to transform the image

- 6 PCT

Gender Bias in Image Embeddings [1]

but ...

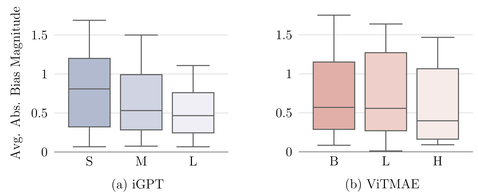
- 3 PCT

CIFAR10 Acc. [1]

[1] averaged across models, and pre–training phases

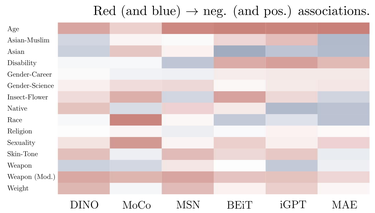
Scaling can help to mitigate social biases

1. Increasing model size leads to a reduction in the average magnitude of biases. → Scaling is a practical strategy to mitigate social biases.
2. But direction of biases remains relatively unchanged. → Models with similar architectures learn similar biases.
3. Inputs resolution and patch size have no systematic effect.



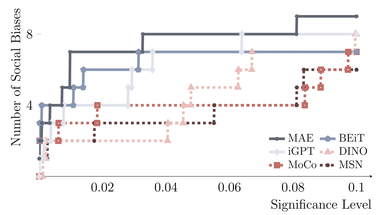
(a) iGPT (b) ViTMAE

Training objectives shape learned social biases



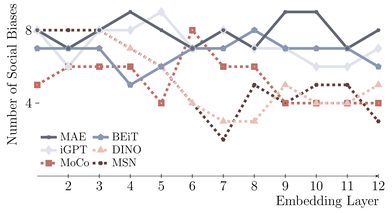
Variability in Learned Social Biases

- To our surprise, models with similar architecture can exhibit opposite social biases despite being trained on the same datasets
- But a small set of consistent biases emerges



Bias Mitigation using Discriminative Training

- On average, discriminative training results in fewer biases than generative training



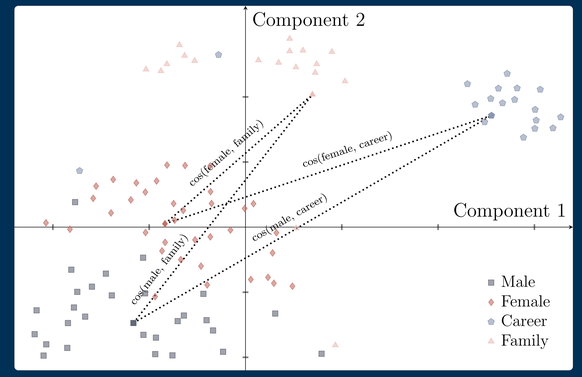
Depth-Dependent Bias Manifestation

- Intensity of biases differs between layers, likely due to increasing semantic interpretability
- Disparity in biases between generative and discriminative models widens in the deeper layers

What’s next?

- To mitigate bias, there are many variables with moderate impact (no one-size-fits-all solution)
- We should explore architectural modifications that achieve significant fairness gains without compromising performance (see e.g., SoLU activation functions for interpretability)

Social biases in vision transformers are **not just a result of training data** but are affected by model design and training objectives.



Gender bias in image embedding from ViTMAE: t-SNE (n=2) reveals that “female” is more closely associated with “family” rather than “career”, whereas “male” has a comparable association with both attributes.

UNIVERSITY OF MANNHEIM