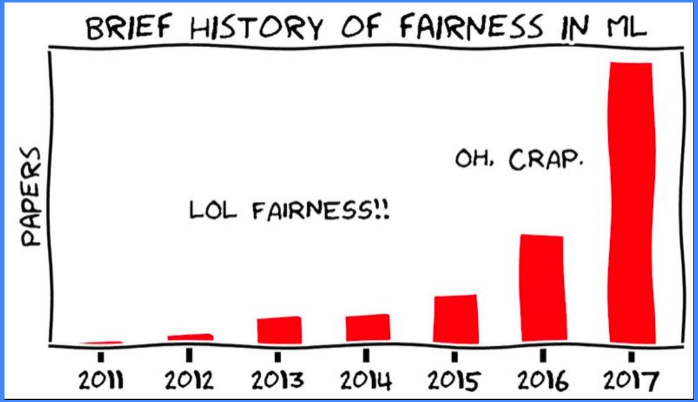
Universiteit van Amsterdam

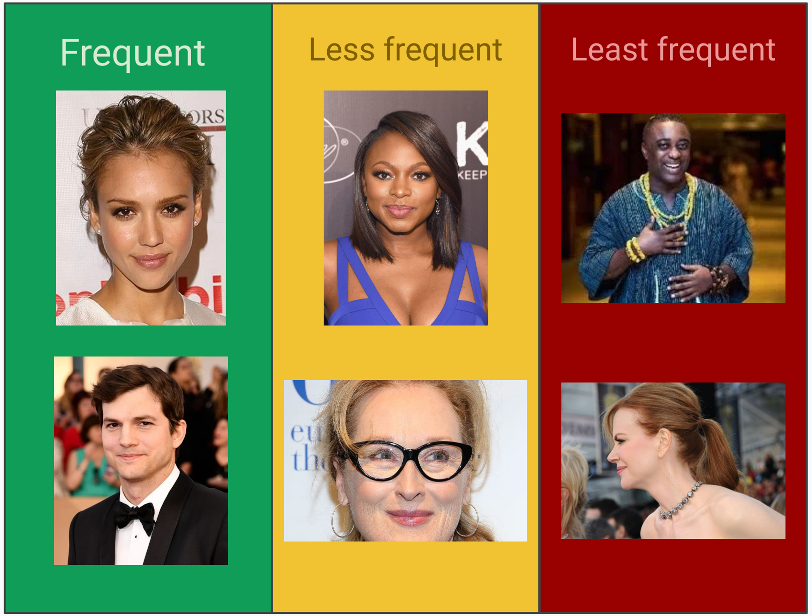
Fairness in Facial Recognition



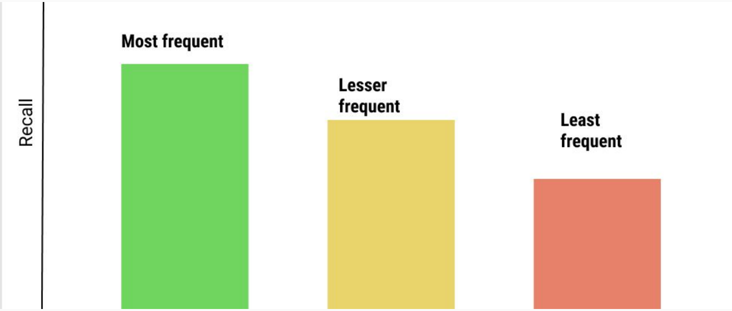
Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure

Dante Niewenhuis, Lucas Fijen, Jonathan Mitnik and Pieter de Marez Oyens

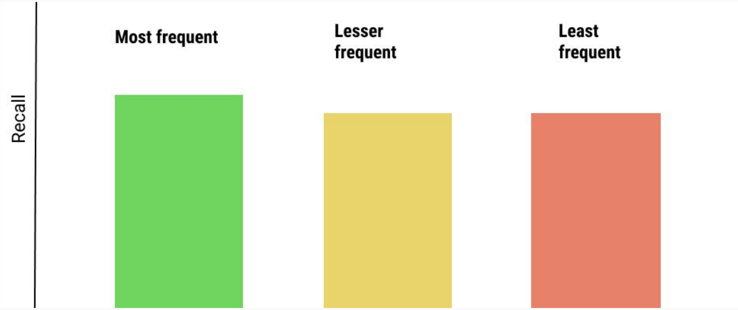
Original Paper’s Goal: Are these **faces?**



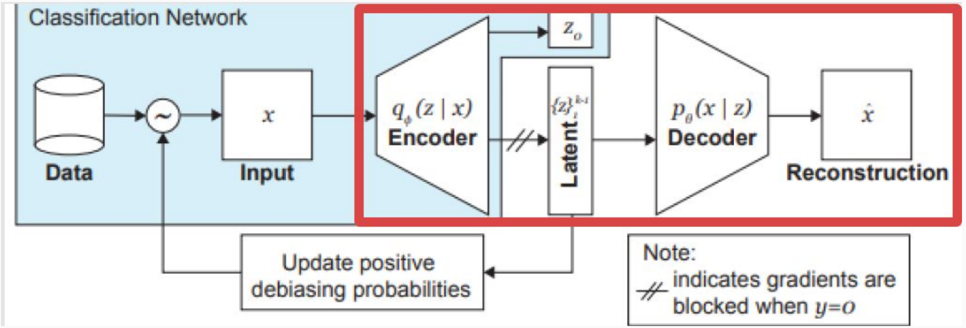
What **Biased** looks like



What **Debiased** *ideally* looks like



Original Paper: VAE and Classification



Input: Image (64-by-64)

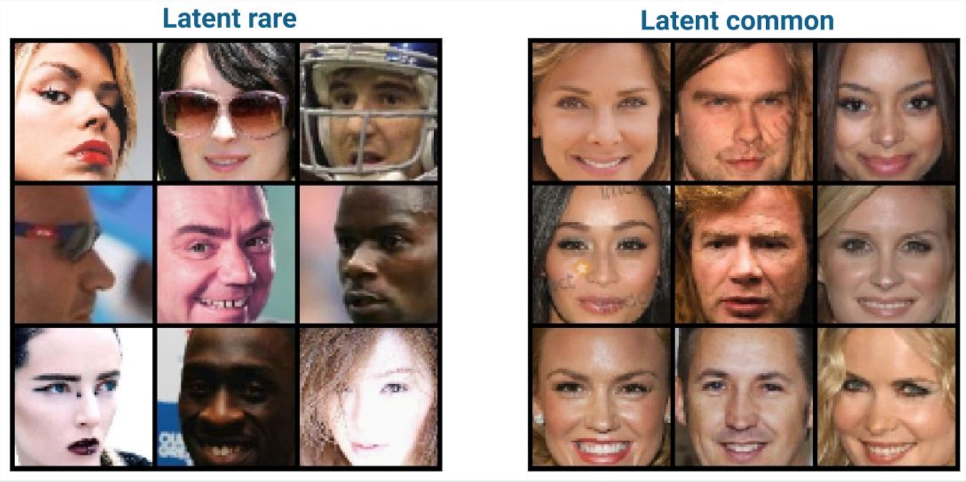
VAE: Encodes and Decodes an image

- Encoding
  - Connects to **classification layer** and the **latent space**
- Goal: To train the encoder-decoder optimally to recognize faces.

Original Paper: Debiasing by Sampling ⅕

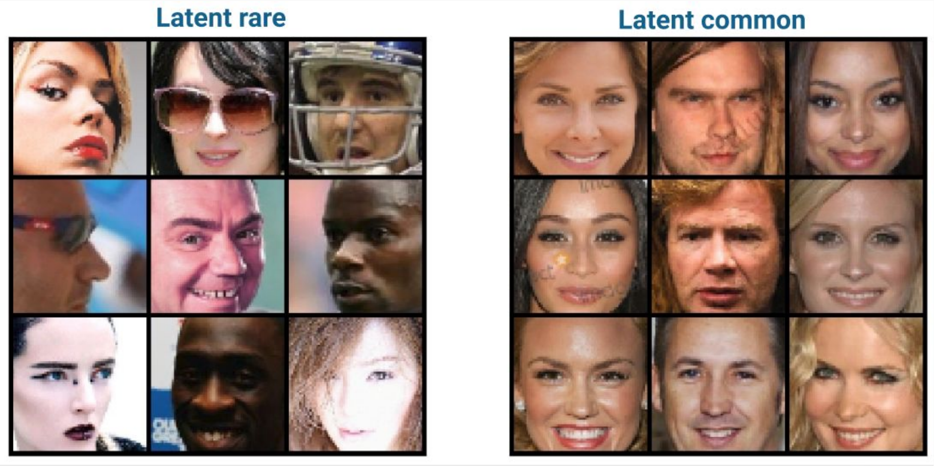
GOAL:
Create a weight for each image in the training set based on latent rarity

=> Sample more from ***latent rare*** over *latent common*



Original Paper: Debiasing by Sampling ⅖

Where to go:
Calculate Q(z|X) which describes how *common* a latent space is for images.



Original Paper: Debiasing by Sampling ⅗

Calculate Q using **Q_i**

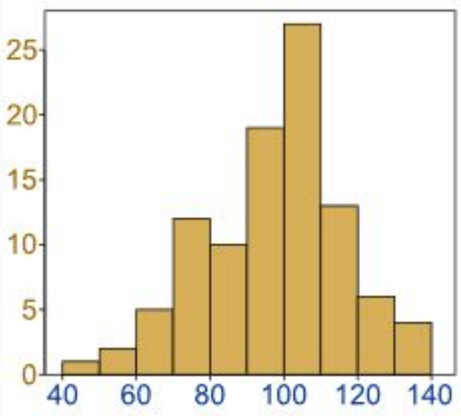
$$\hat{Q}(z|X) \propto \prod_i \hat{Q}_i(z_i|X)$$

Assumes each latent variable **Q_i** is independent

Original Paper: Debiasing by Sampling 4/5

Calculate Q_i by *frequency*

Each epoch, build up histogram of *every Q_i* for the entire dataset **X**



Original Paper: Debiasing by Sampling

Aggregate Q_i to form W
Get sampling probabilities of images

$$\mathcal{W}(z(\boldsymbol{x})|X) \propto \prod_i \frac{1}{\hat{Q}_i(z_i(\boldsymbol{x})|X) + \alpha}$$

=> *Latent **rare** images have **higher W***

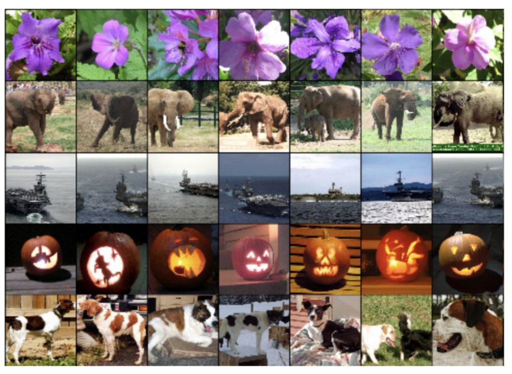
❗ The *alpha* parameter measures how much of the debiasing is applied. A lower alpha means more debiasing.

Universiteit van Amsterdam

Original Paper: Experiments

Training

Faces
from **CelebA**



Non-faces
from **Imagenet**

Evaluation: PPB

Images annotated between
skin-colors lightest to darkest

Original Paper: Evaluation method on PPB

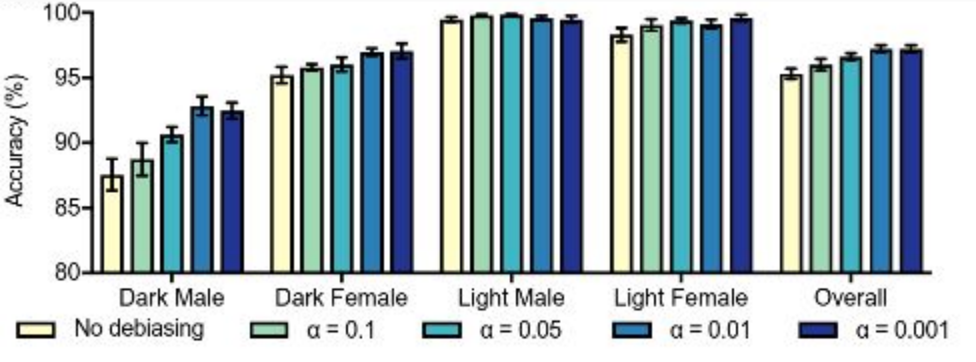
1. Evaluate on faces ***only***

2. For each image, split image into **sub-images**

3. Find at least **one face** in the sub-images to classify the original image as true.

4. Measure **“accuracy”** by the number of sub-images containing at least one face.

Original Paper: Results



	$\mathbb{E}[\mathcal{A}]$ (Precision)	$Var[\mathcal{A}]$ (Measure of Bias)
No Debiasing	95.13	28.84
$\alpha = 0.1$	95.84	25.43
$\alpha = 0.05$	96.47	18.08
$\alpha = 0.01$	97.13	9.49
$\alpha = 0.001$	**97.36**	**9.43**

Our implementation: Problems

- Unstable calculations
- Bad evaluation method

UNIVERSITEIT VAN AMSTERDAM

Our implementation: Unstable calculations

- Formula for the weights is unstable when using a high-dimensional latent space
- W will either underflow or overflow
- This makes the formula unusable

$$\mathcal{W}(z(\mathbf{x})|X) \propto \prod_i \frac{1}{\hat{Q}_i(z_i(\mathbf{x})|X) + \alpha}$$

Universiteit van Amsterdam

Our implementation: Alternative formulas

- Alternate function was found in the code of one of the authors[1]
- Instead of the product the max is taken
- Other methods were tried but were not so consistent

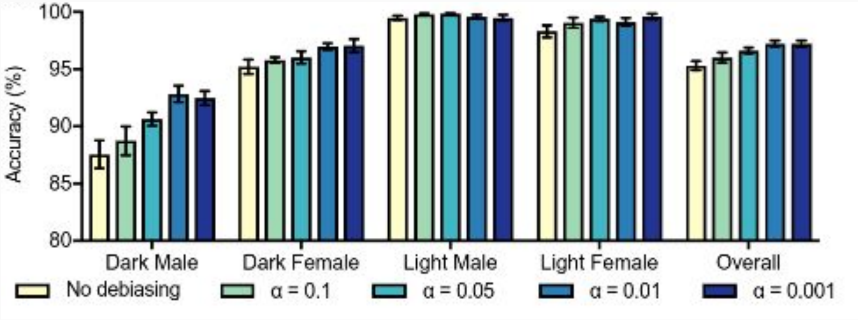
$$\mathcal{W}\left(z\left(\mathbf{x}\right)|X\right) = \max_{i}\left(\frac{1}{\hat{Q}_i\left(z|X\right) + \alpha}\right)$$

1. https://github.com/aamini/introtodeeplearning/blob/master/lab2/solutions/Part2_Debiasing_Solution.ipynb

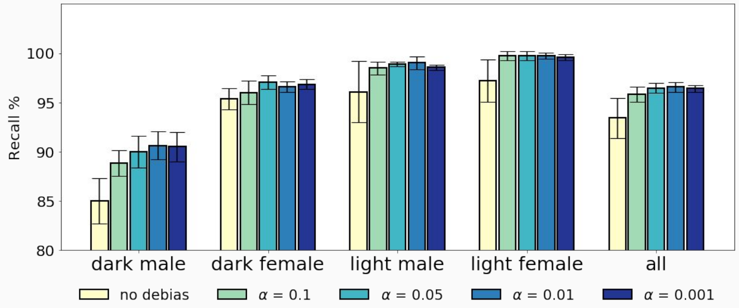
Universiteit van Amsterdam

Our implementation: Results

Using max debias method similar results were achieved



Original Results



Our Results

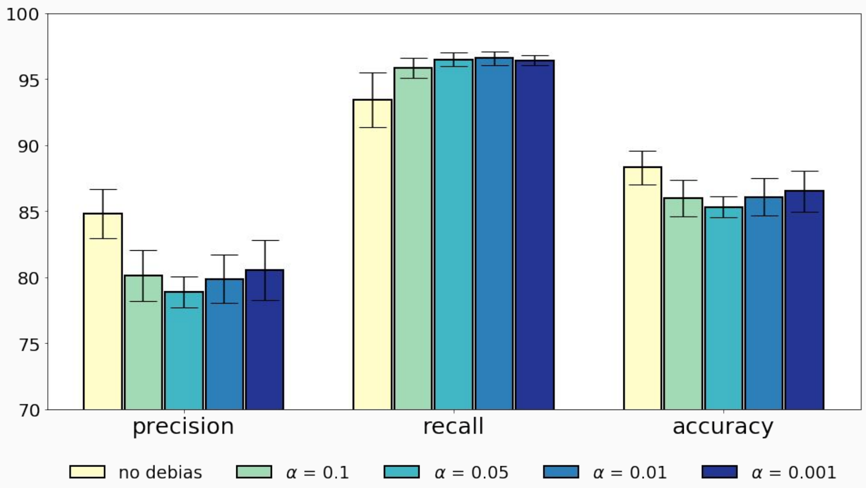
Our implementation: Extending evaluation

- Original evaluation **only** focuses on classifying faces correctly
  - This is disregarding **half of the problem** in our opinion
- We have extended the evaluation to also classify non-faces
- Non-faces are evaluated using the same method
- Now we can calculate **recall**, **precision** and **accuracy**

Universiteit van Amsterdam

Our implementation: Extending results

- The extended results shows the bigger picture
- Three conclusion can be made
  - Recall **increases**
  - Precision **decreases stronger**
  - Accuracy **decreases**



Precision, Recall and accuracy of models with varying degrees of debiasing

Conclusion

Pros:

- Results were able to be **reproduced**
- The **variance in recall** across groups **reduces with debiasing**
- **Increased recall** for all classes

Cons:

- Drop in **precision** and **accuracy**
- Metric **accuracy** was not used correctly, which can be **misleading**
- Evaluation dataset contains **great contrast** in image quality

Final verdict

The paper itself was **not sufficient** for reproduction,

however, the unofficial code **filled the critically missing** gaps.

Universiteit van Amsterdam

Thank you for listening