Bias in Face Detection (?)

Re-implementation and extension of the Paper:
Uncovering and Mitigating Algorithmic Bias Through Learned Latent Structure

by Frederic Chamot, Maximilian Knaller, Luisa Ebner and Julio López González

FACT-AI Jan 2020

Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

The New York Times

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

Facial Recognition's Many Controversies, From Stadium Surveillance to Racist Software

When the Robot Doesn't See Dark Skin

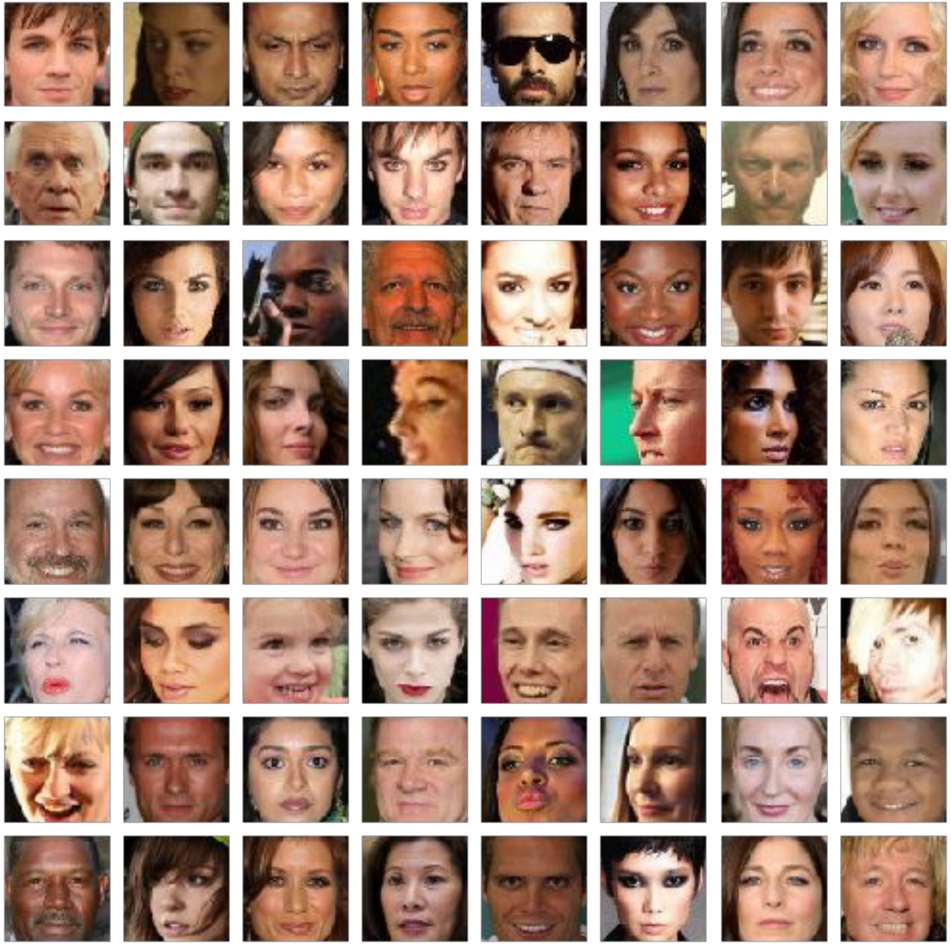
By Joy Buolamwini

Ms. Buolamwini is the founder of the Algorithmic Justice League.

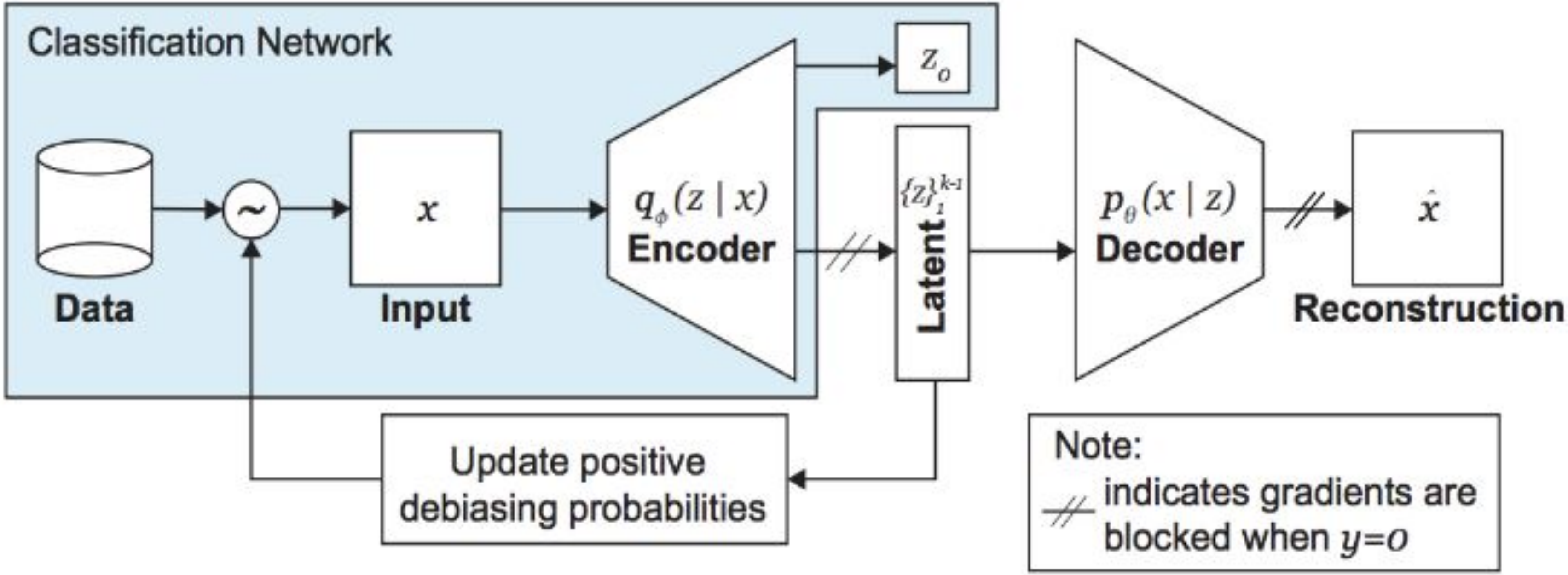
June 21, 2018

Gender and racial bias found in Amazon's facial recognition technology (again)

Research shows that Amazon's tech has a harder time identifying gender in darker-skinned and female faces



Debiasing VAE - Intuition

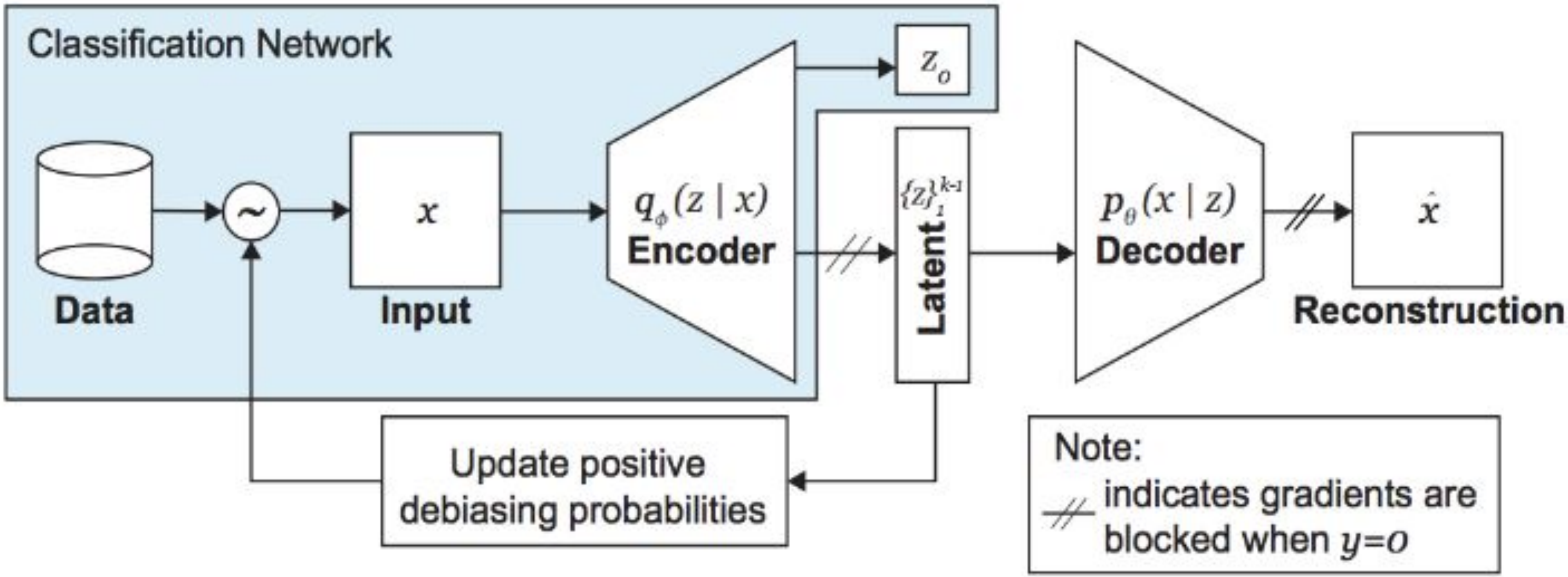


Debiasing VAE - Loss

$$\mathcal{L}_{SAMPLE} = c_{class} \cdot \underbrace{\left[y \log\left(\frac{1}{\hat{y}}\right) \right]}_{\mathcal{L}class(y,\hat{y})} + y \cdot \underbrace{\left[c_{recon} \cdot \underbrace{\left[\|x - \hat{x}\|_2 \right]}_{\mathcal{L}_{recon}(x,\hat{x})} + c_{KL} \cdot \underbrace{\left[\frac{1}{2} \sum_{j=0}^{k-1} \left(\sigma_j + \mu_j^2 - 1 - \log(\sigma_j) \right) \right]}_{\mathcal{L}_{KL}(\mu,\sigma)} \right]}_{\mathcal{L}_{VAE}}$$

$c_{class} = 0.2$ $c_{recon} = 0.1$ $c_{KL} = 0.05$

Debiasing VAE - Intuition



Debiasing VAE - Sample Probabilities

1. Compute a histogram $\hat{Q}_i\left(z_i|X\right)$ on each latent dim
2. Compute latent probability of samples $\mathcal{P}_{Q_i}\left(X|\hat{Q}_i(z_i|X)\right)$ by the probability mass of their respective bins
3. Compute sample weights with smoothing factor $\boldsymbol{\alpha}$

$$\mathcal{W}(X|Z) = \sum_i \log\left(\frac{1}{\mathcal{P}_{Q_i}\left(X|\hat{Q}_i(z_i|X)\right) + \alpha} + 1\right)$$

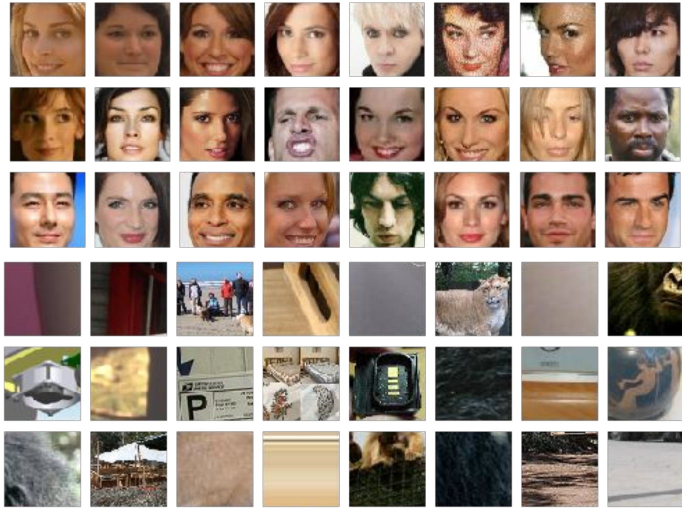
4. Compute final sample probabilities

$$\mathcal{P}(X|Z) = \frac{\mathcal{W}(X|Z)}{\sum_j \mathcal{W}(x_j|Z)}$$

Datasets - Training

CelebA + ImageNet (MIT)

x ~ 110 K



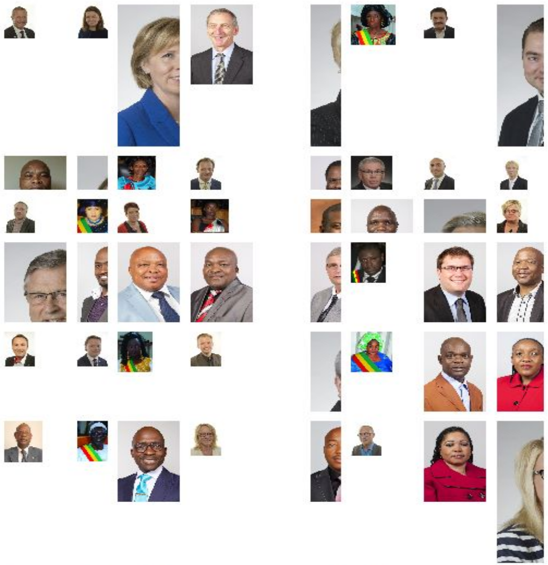
FairFace + OpenImages (OUR)

x ~178 K

Datasets - Bias Evaluation

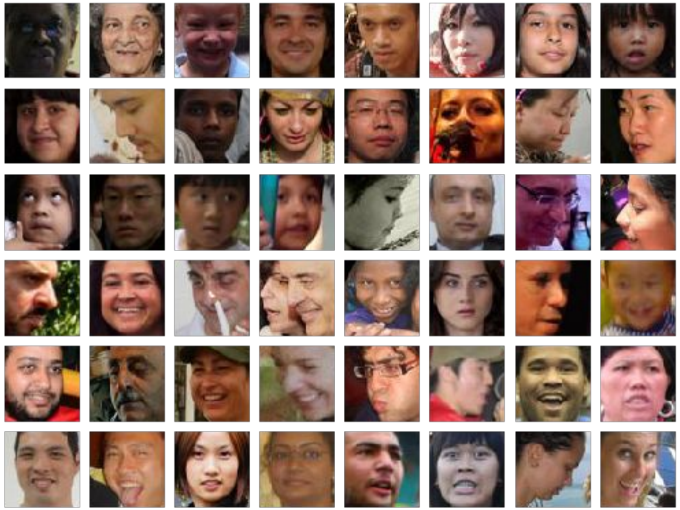
Pilot Parliaments Benchmark (PPB)

x ~1.3 K



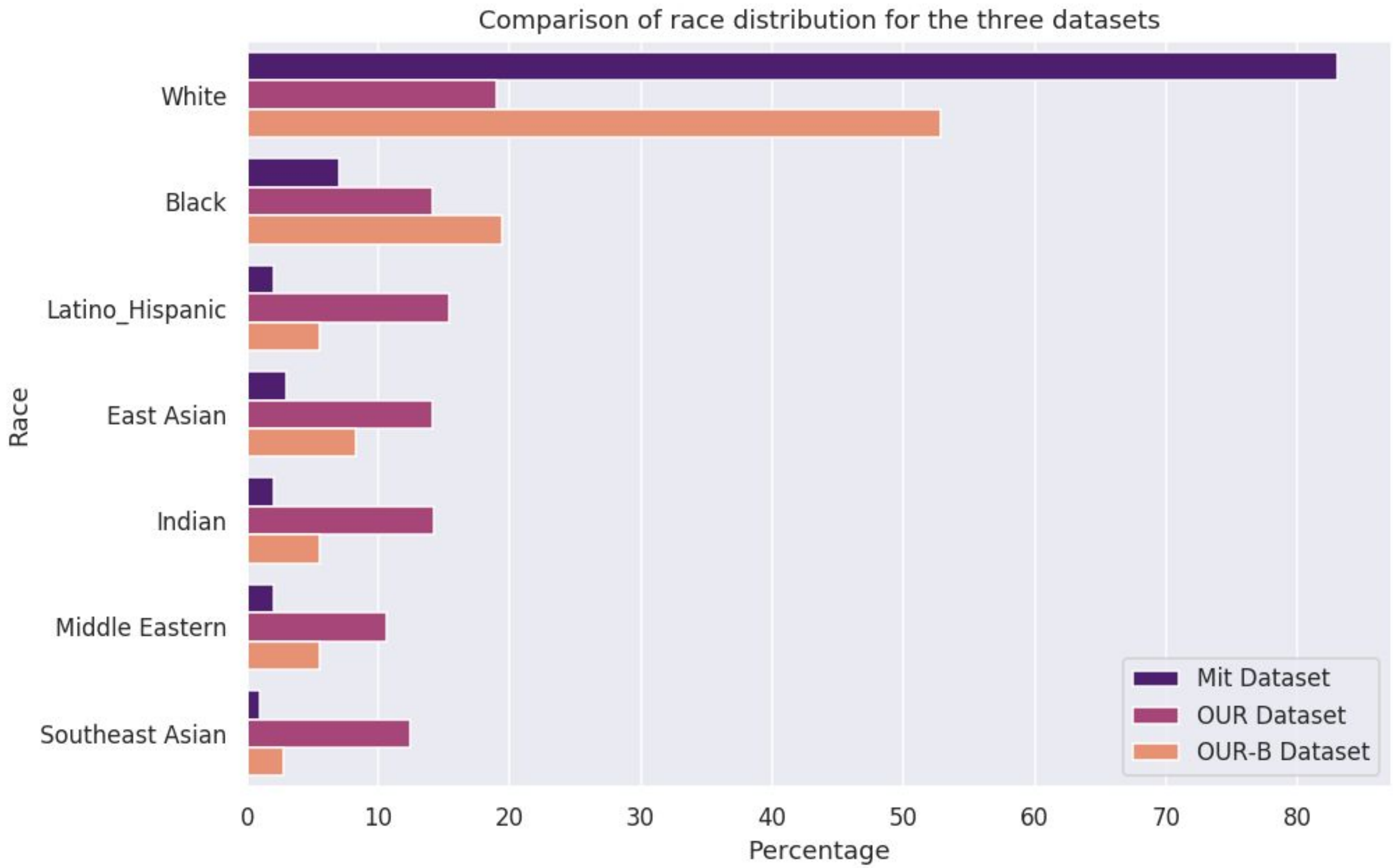
OUR-VAL

x ~22 K



Our Unbiased Reproducible (OUR) Dataset

+ **unbiased** with regard to race and gender features
+ greater in-class diversity / more natural image settings
+ **reproducible** by transparent creation (minimizing faces in counterexamples)
+ detailed annotations (gender/age/race) relevant for bias research for training and bias-validation
+ larger in size
+ training and bias-evaluation use same preprocessing -> no sliding window
+ Can simulate the race distribution bias in MIT

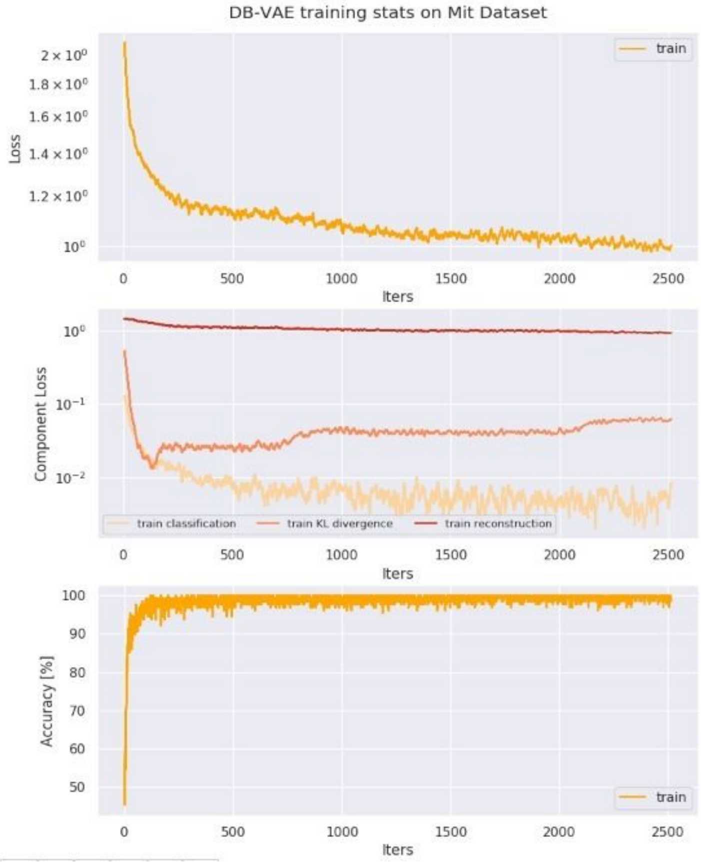


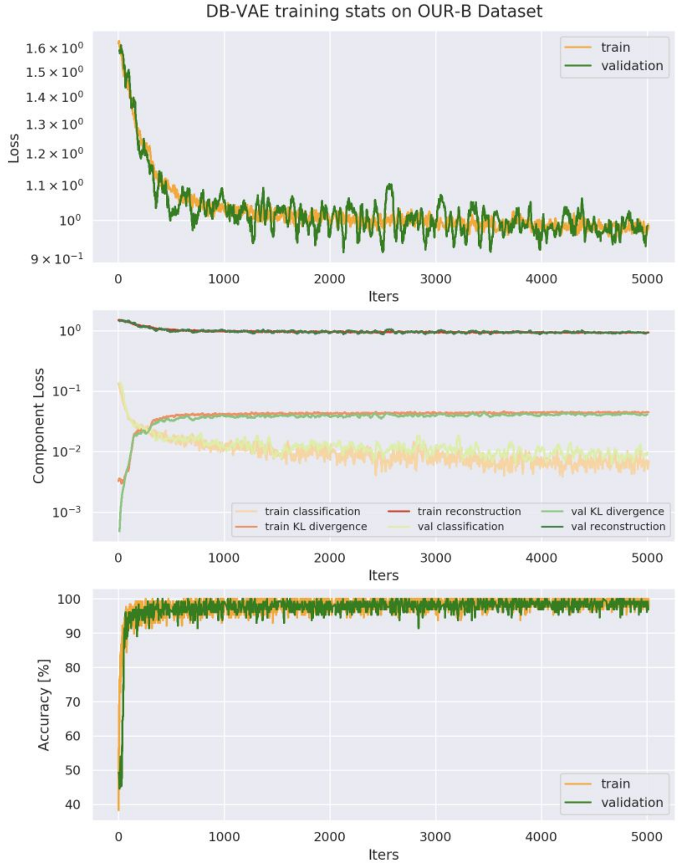
Comparison of race distribution for the three datasets
Race
White
Black
Latino_Hispanic
East Asian
Indian
Middle Eastern
Southeast Asian
0
10
20
30
40
50
60
70
80
Percentage
Mit Dataset
OUR Dataset
OUR-B Dataset

Key contributions:

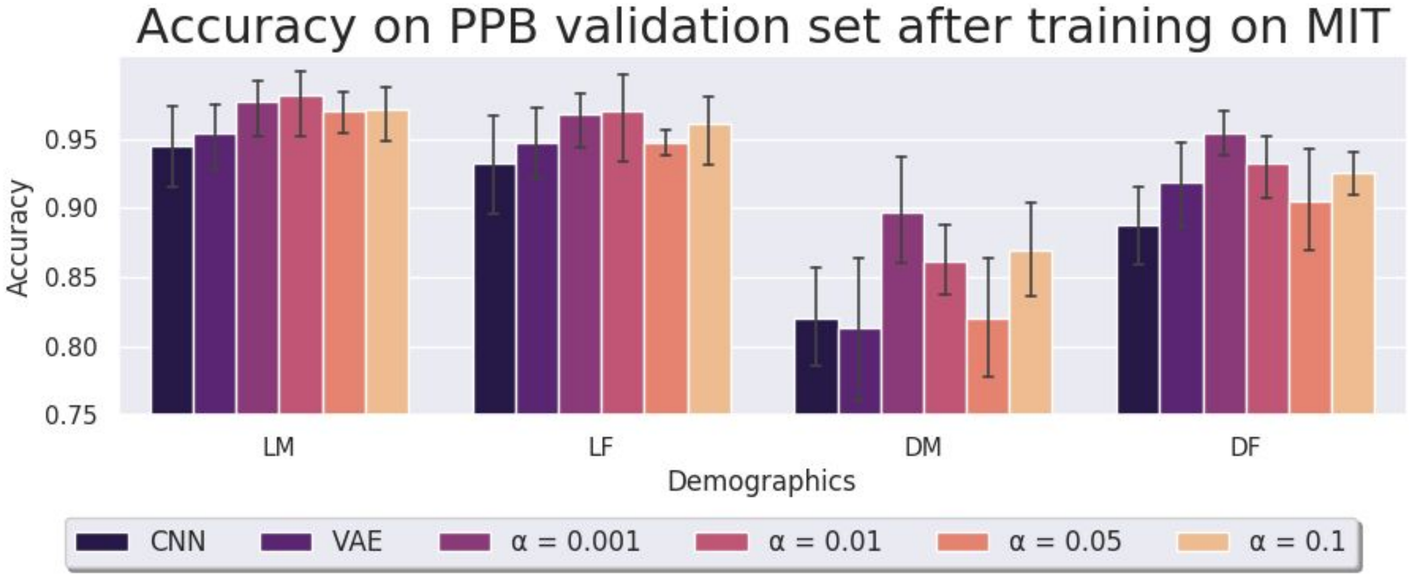
1. Exact replica of the results of MIT paper.
2. In-Depth Reevaluation on a second larger and more versatile dataset.

Results and Discussion

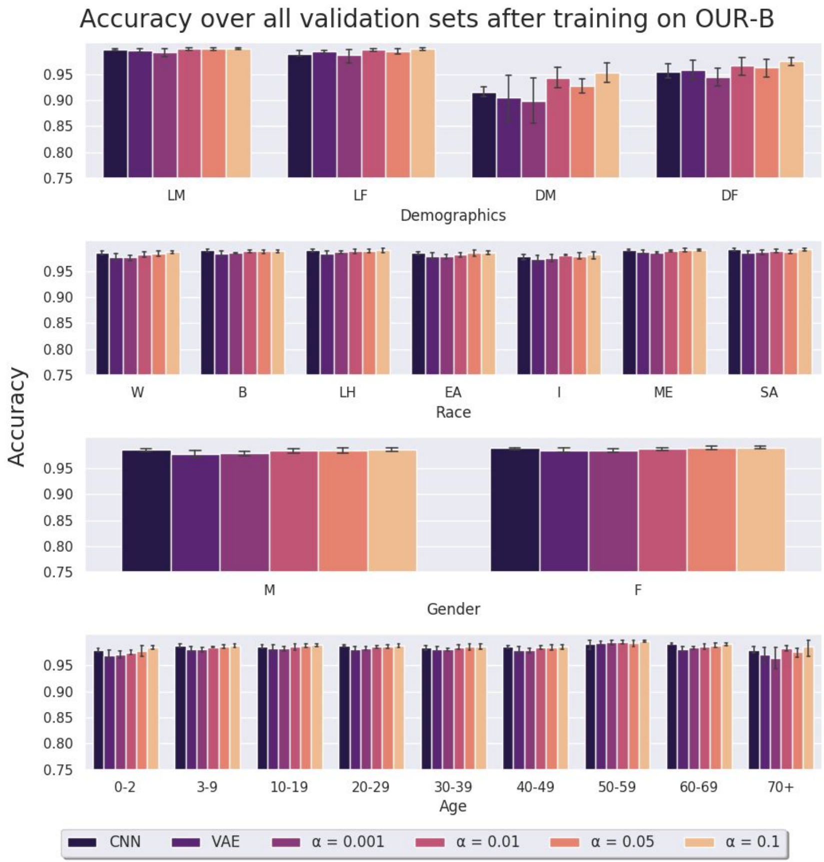




Results - MIT training data



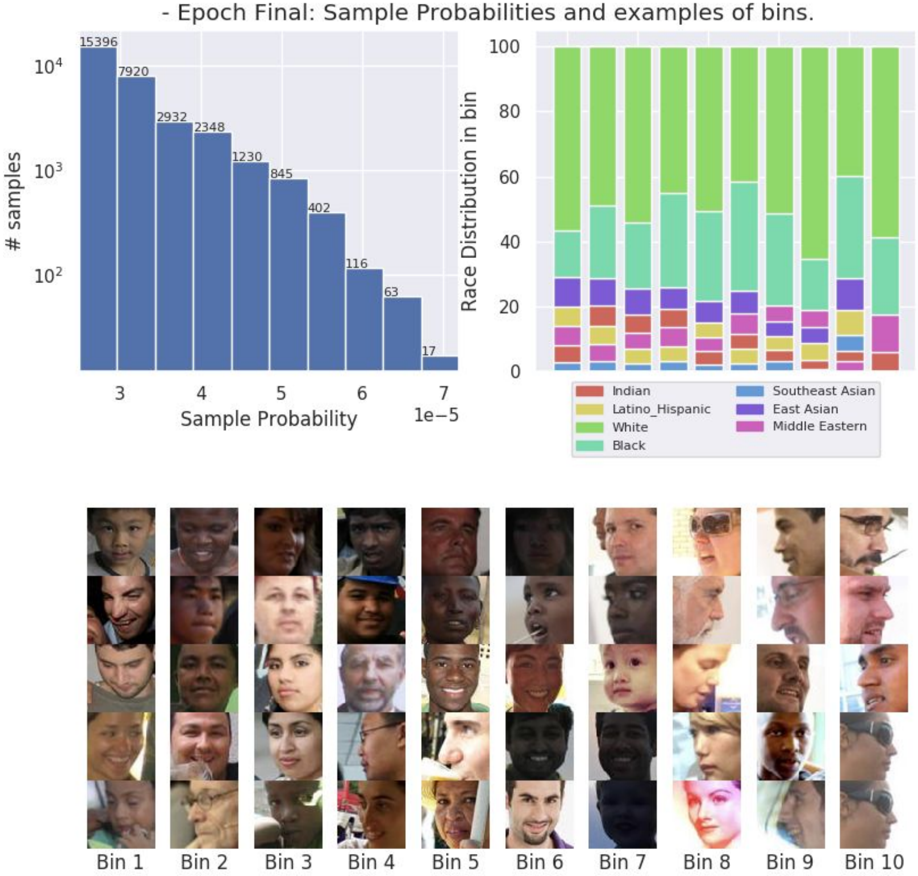
Results - OUR-B training data



Key observations

- + effect not as pronounced as expected
- + accuracy of age/gender unaffected by race bias
- + OUR validation set associated with lower variance

Sample Probabilities



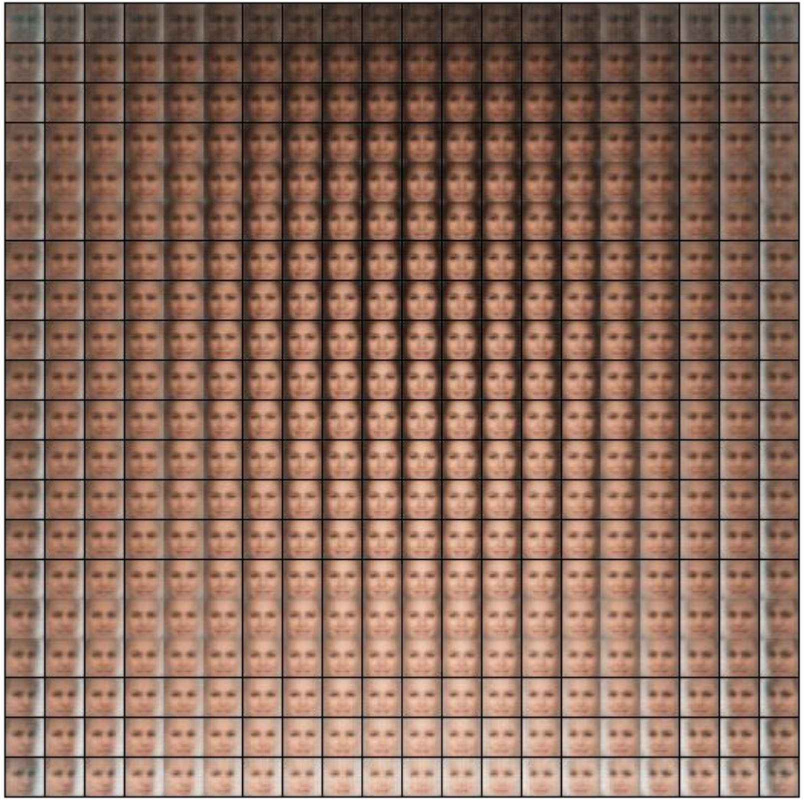
Problem:

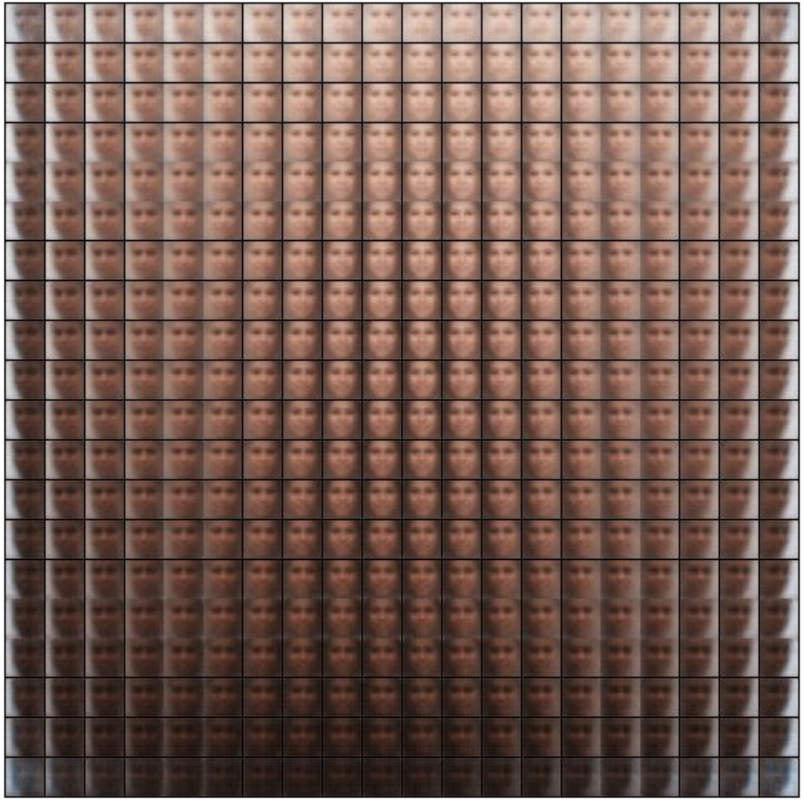
learned latent features

≠

sensitive target features

Latent Space Manifolds





Decoder Reconstructions

Take Home Message

The DB-VAE

+ learns **semantically meaningful** features
+ adapts resampling to **explore underrepresented feature space**

However

- **sensitive features** are learned unsupervised -> difficult to target specifically

Questions???