Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure

Presenters: Madhumita Sakthi
Pratyush Kar

Problem setup

- Face detection system has low accuracy for dark women between 18-30 years of age (Klare et al.)
- Mitigate bias in the training data - model training - unsupervised

Random Batch Sampling During Standard Face Detection Training

Homogenous skin color, pose

Mean Sample Prob: 7.57×10^{-6}

Batch Sampling During Training with Learned Debiaising

Diverse skin color, pose, illumination

Mean Sample Prob: 1.03×10^{-4}

So far..

- **Resampling for class imbalance:**
  - Address class imbalance, as opposed to biases in individual classes
  - Duplicating minority classes for mitigating class imbalance- but does not run adaptively
  - This also required manual annotation of classes
- **Generating debiasing data:**
  - Generate fair training data
  - Preprocessing data transformations that mitigate discrimination
  - Rely on artificially generated dataset
- **Clustering to identify bias:**
  - *K-means* clustering to identify clusters in input data prior to training and inform resampling
  - Does not scale to high dimensional dataset

Problem setup

- Binary classification setup
- Use VAEs to learn the latent structure
  - Re-weight the importance of certain data points while training
- Learn the functional mapping *f(x)*

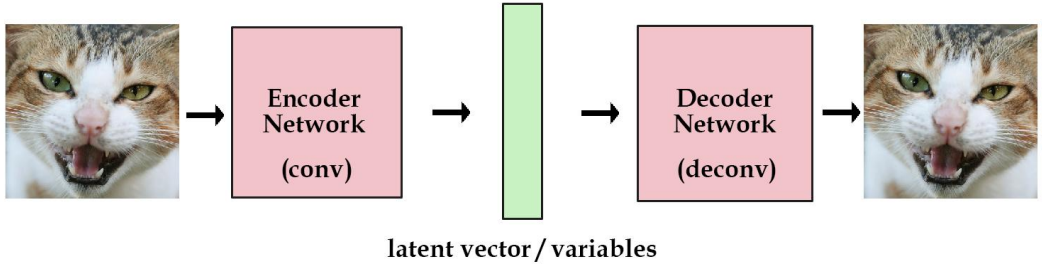
$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta})$$

- Fair classifier:

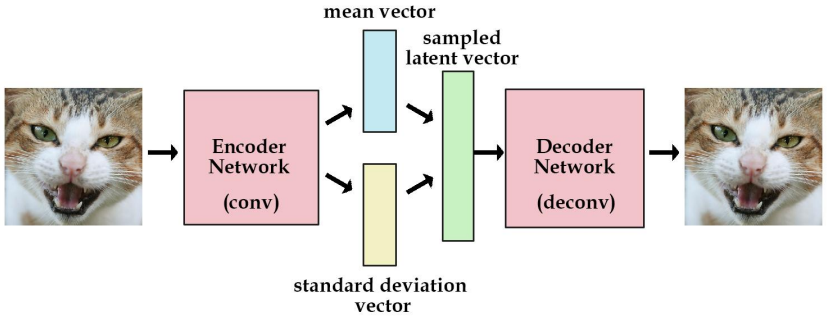
Definition 1 *A classifier,* $f_{\boldsymbol{\theta}}(x)$*, is* ***biased*** *if its decision changes after being exposed to additional sensitive feature inputs. In other words, a classifier is fair with respect to a set of latent features,* z*, if:* $f_{\boldsymbol{\theta}}(x) = f_{\boldsymbol{\theta}}(x, z)$

- Where, *z* could be skin color, gender, age

AE vs. VAEs

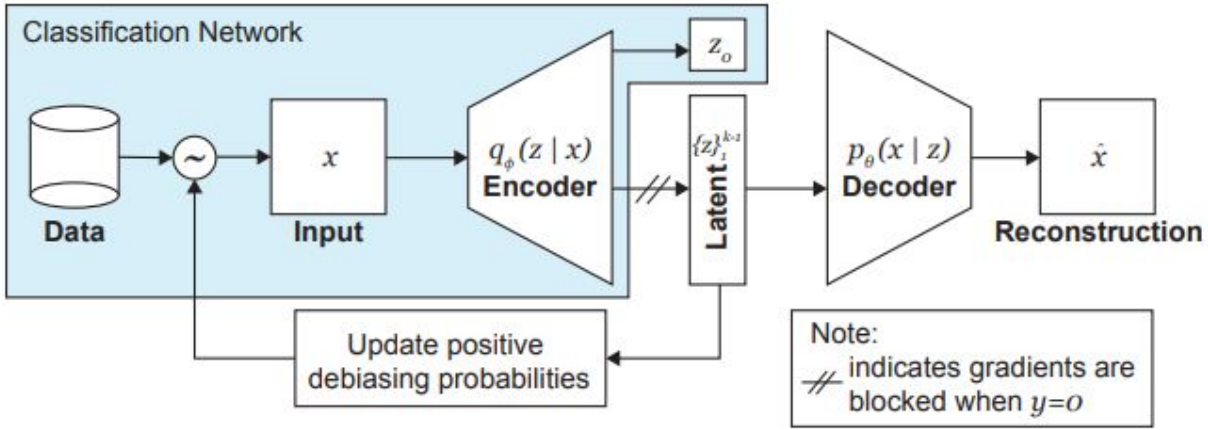


- Learn latent vectors that approximates unit gaussian
- Minimize: KL-divergence + MSE
- Reparameterization trick :



PC: http://kvfrans.com/variational-autoencoders-explained/

Network Architecture



Loss functions

- *Ly(y,y')* is the supervised loss (cross-entropy)
- *Lx(x,x')* is the reconstruction loss (L-p norm)
- *LKL(u,s)* is the KL divergence loss

$$\mathcal{L}_{TOTAL} = c_1 \underbrace{\left[\sum_{i \in \{0,1\}} y_i \log\left(\frac{1}{\hat{y}_i}\right) \right]}_{\mathcal{L}_y(y,\hat{y})} + c_2 \underbrace{\left[\|x - \hat{x}\|_p \right]}_{\mathcal{L}_x(x,\hat{x})}$$
$$+ c_3 \underbrace{\left[\frac{1}{2} \sum_{j=0}^{k-1} (\sigma_j + \mu_j^2 - 1 - \log(\sigma_j)) \right]}_{\mathcal{L}_{KL}(\mu,\sigma)} \qquad (2)$$

- *Negative sample's* gradients from decoder and latent space are not backpropagated

Algorithm for Automated debiasing

Algorithm 1 Adaptive re-sampling for automated debiasing of the DB-VAE architecture

Require: Training data $\{X, Y\}$, batch size b

Initialize weights $\{\phi, \theta\}$
for each epoch, E_t **do**
  Sample $z \sim q_\phi(z|X)$
  Update $\hat{\mathcal{Q}}_i(z_i(x)|X)$
  $\mathcal{W}(z(x)|X) \leftarrow \prod_i \frac{1}{\hat{\mathcal{Q}}_i(z_i(x)|X)+\alpha}$
  while $iter < \frac{n}{b}$ **do**
    Sample $\boldsymbol{x}_{batch} \sim \mathcal{W}(z(x)|X)$
    $L(\phi, \theta) \leftarrow \frac{1}{b} \sum_{i \in \boldsymbol{x}_{batch}} \mathcal{L}_i(\phi, \theta)$
    Update: $[w \leftarrow w - \eta \nabla_{\phi,\theta} \mathcal{L}(\phi, \theta)]_{w \in \{\phi,\theta\}}$
  end while
end for

Experiments

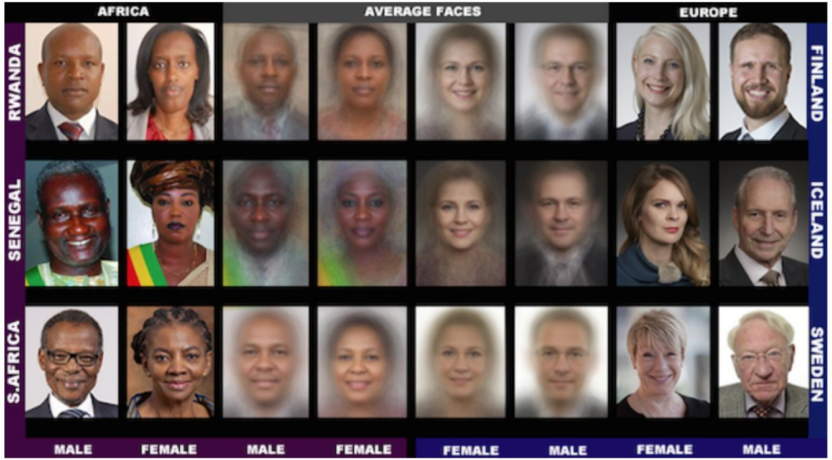
- Facial detection problem
- Binary classification problem to identify face in an image
- Positive examples: Contains faces
- Negative examples: Does not have a face
- Train a full DB-VAE to learn latent structure of **positive examples**
- For **negative examples** de-biased sampling is not done

Dataset (Face Detection)

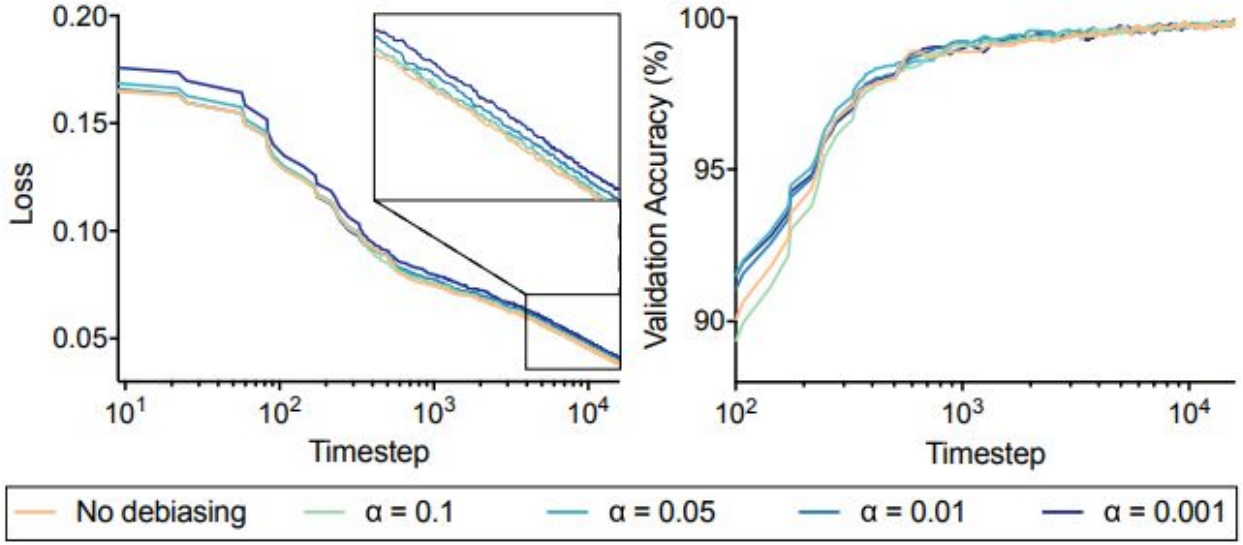
- Contains 4 x 10^5 images with an equal number of positive and negative examples
- Each image of size 64 x 64
- Positive examples were taken from CelebA dataset
- Negative examples taken from wide-variety of non-human classes in ImageNet

Evaluation

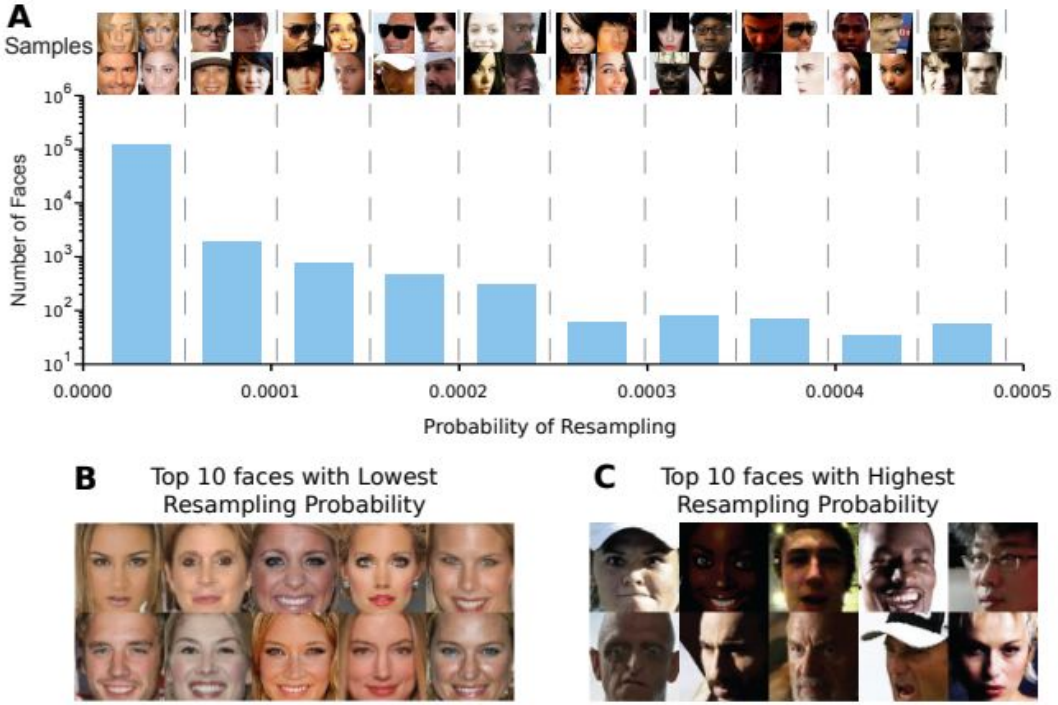
- Pilot Parliaments Benchmark (PPB) Dataset
  - 1270 photos of male and female parliamentarians
  - Images consistent in pose, illumination and facial expression
  - Each image is annotated with sex-based labels (‘male’ and ‘female’)
  - Each image is also labelled with the skin tone (‘lighter’ and ‘darker’)
  - Bias is evaluated by calculating the variance in the classifier accuracy for these 4 intersectional classes



Loss Evolution and Validation Accuracy



Sampling Probabilities Over Training Set



Decreased Categorical Bias With DB-VAE

- DB-VAE improved accuracy on dark males but never reached the accuracy of white males- **inherent lack of data**
- Usage of DB-VAE did not decrease the performance for white males

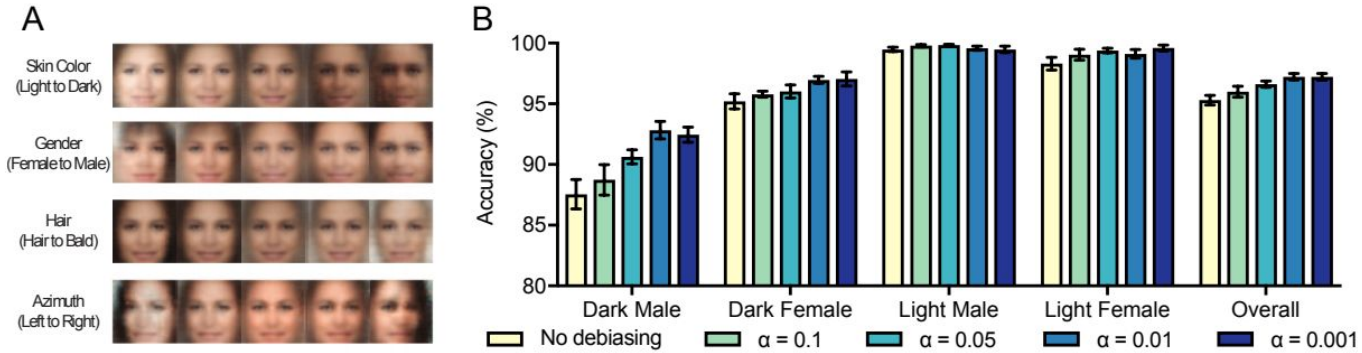


Figure 5: **Increased performance and decreased categorical bias with DB-VAE.** The model learns latent features such as skin color, gender, hair (**A**) and demonstrates increased performance and decreased categorical bias with learned debiasing (**B**).

Accuracy and Bias on PPB Dataset

- Overall precision increases with increasing debiasing power (decreased alpha)
- Decreased variance in accuracy with increasing debiasing power

Table 1: **Accuracy and bias on PPB test dataset.**

	$\mathbb{E}[\mathcal{A}]$ (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**

Questions?