Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure

(Amini & Soleimany et al. 2019)

FACT ‘20

By: Puck de Haan, Paul ten Kaate, Tim van Loenhout, Jan Erik van Woerden

Content

❖ Introduction
❖ Method (principles)
❖ Issues
❖ Results
❖ Discussion and conclusion

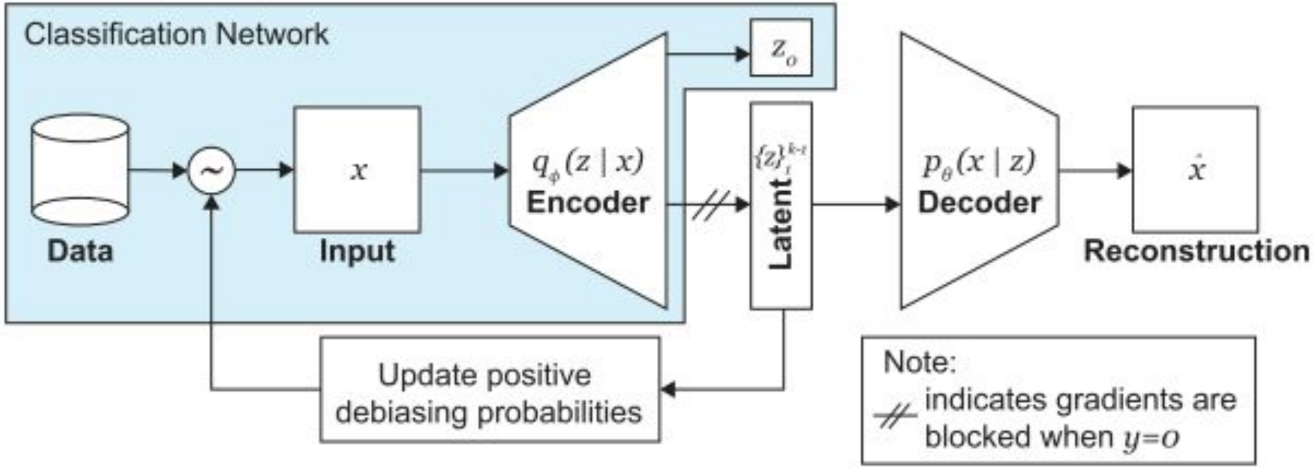
Fairness

- Classification task (f).
- Notion of fairness:
  - Classifier should be unbiased w.r.t. latent variables.
  - Every subset of data treated the same.
- Facial recognition subject to algorithmic bias.
  - Skewed training data (x)
  - Gender, skin tone, age, etc.

Approach

- Solution: Uncovering bias in dataset.
- Combine VAE and Classifier
  - Using latent space of data (z)
- Adjust sampling probabilities
- Straightening dataset

Method -- Model



Source: (Amini & Soleimany et al. 2019)

Method -- Adaptive Resampling

- Proposed in paper:

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 $\hat{Q}_i(z_i(x)|X)$

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

  while $iter < \frac{n}{b}$ **do**

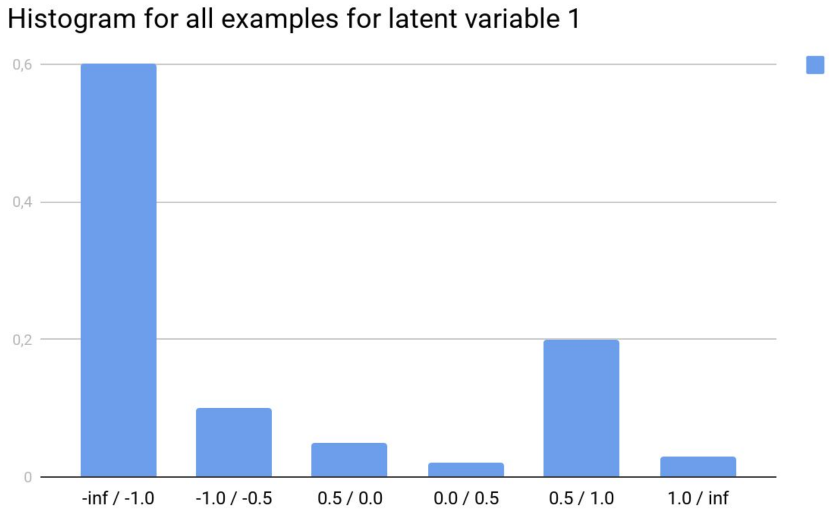
    Sample $\mathbf{x}_{batch} \sim \mathcal{W}(z(x)|X)$

    $L(\phi, \theta) \leftarrow \frac{1}{b} \sum_{i \in \mathbf{x}_{batch}} \mathcal{L}_i(\phi, \theta)$

    Update: $[w \leftarrow w - \eta \nabla_{\phi,\theta} \mathcal{L}(\phi, \theta)]_{w \in \{\phi, \theta\}}$

  end while

end for



Implementation issues

- Adaptive resampling
- Model architecture
  - Activations
  - Padding
- Loss
  - Weights
  - Reconstruction loss
- Latent dimension

$$\mathcal{L}_{TOTAL} = c_1 \underbrace{\left[\sum_{i \in \{0,1\}} y_i \log\left(\frac{1}{\hat{y}_i}\right) \right]}_{\mathcal{L}_y(y,\hat{y})} + c_2 \underbrace{\left[\|x - \hat{x}\|_p \right]}_{\mathcal{L}_x(x,\hat{x})}$$

$$+ c_3 \underbrace{\left[\frac{1}{2} \sum_{j=0}^{k-1} (\sigma_j + \mu_j^2 - 1 - \log(\sigma_j)) \right]}_{\mathcal{L}_{KL}(\mu,\sigma)}$$

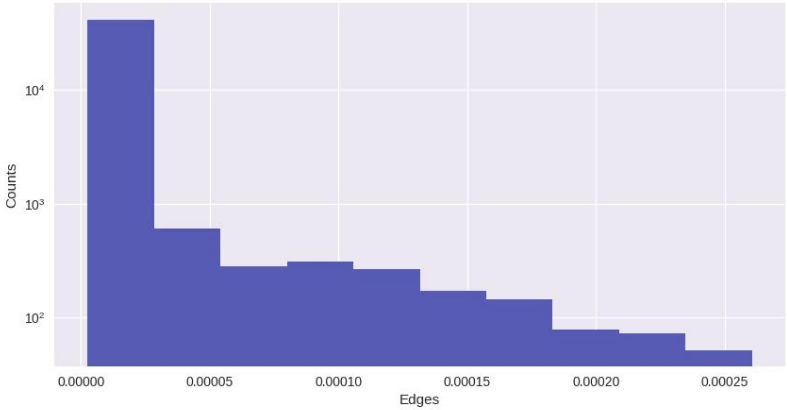
Complete loss function of the model.

Results -- Adaptive Resampling

- Frequent → low probability
- Rare → high probability

Faces with the highest sampling probability.

Faces with the lowest sampling probability.

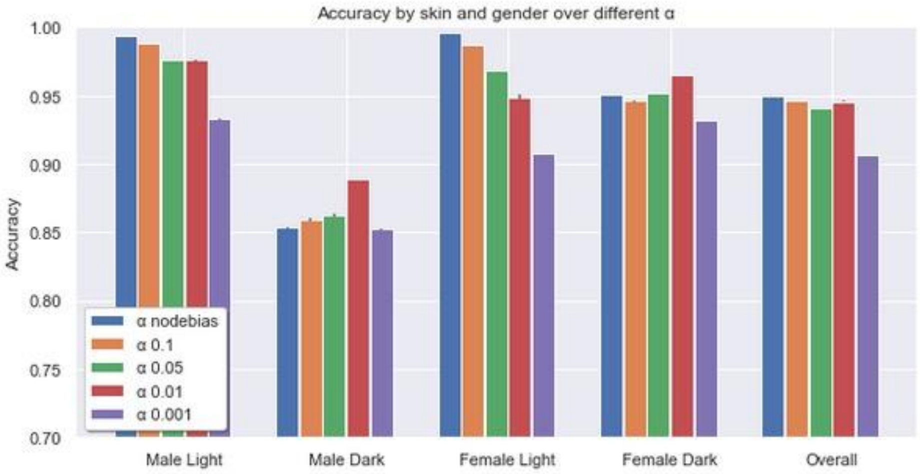


Sampling probabilities of training data after debiasing.

Results -- Accuracy / Bias

- Test-set
  - Portrait images
  - Genders
  - Skin tones
- Baseline model performs best accuracy-wise.
- Measure of bias decreases when using higher α.

	$\mathbb{E}[\mathcal{A}]$: Recall	$Var[\mathcal{A}]$: Measure of Bias
No debias	94.86	33.65
$\alpha = 0.1$	94.52	27.39
$\alpha = 0.05$	93.97	20.52
$\alpha = 0.01$	94.49	11.35
$\alpha = 0.001$	90.61	10.75



Discussion -- Reproducibility

- Reduced bias
- Decreasing overall accuracy
- Missing and infeasible settings

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

Paper results

	$\mathbb{E}[\mathcal{A}]$: Recall	$Var[\mathcal{A}]$: Measure of Bias
No debias	92.39	46.20
$\alpha = 0.1$	91.78	36.39
$\alpha = 0.05$	90.76	32.33
$\alpha = 0.01$	90.52	22.00
$\alpha = 0.001$	85.22	22.42

Our results

Discussion -- Posterior collapse

Mean over reconstruction loss

Sum over reconstruction loss

Conclusion

- Not 100% reproducible
- Only with severe restrictions

Sources

- Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. 2019. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure (*AIES '19*). Association for Computing Machinery, New York, NY, USA, 289–295.