Flexibly Fair Representation Learning by Disentanglement

Elliot Creager[1 2] David Madras[1 2] Jörn-Henrik Jacobsen[2]
Marissa A. Weis[2 3] Kevin Swersky[4] Toniann Pitassi[1 2] Richard Zemel[1 2]

June 13, 2019

[1]University of Toronto [2]Vector Institute [3]University of Tübingen [4]Google Research

Why Fair Representation Learning?

$$\text{Fair Representation: } [\mathbf{x}, a] \xrightarrow{f} \mathbf{z} \xrightarrow{g_1} \hat{y}_1$$
$$\mathbf{z} \xrightarrow{g_2} \hat{y}_2$$
$$\mathbf{z} \xrightarrow{g_3} \hat{y}_3$$
$$\dots$$

Given sensitive attribute $a \in \{0, 1\}$, we want:

- $\mathbf{z} \perp a$ (demographic parity) with $\mathbf{z} = f(\mathbf{x}, a)$
- $\mathbf{z}$ maintains as much info about $\mathbf{x}$ as possible

A fair representation acts as a group parity bottleneck

Current approaches are flexible w.r.t. downstream task labels (Madras et al., 2018) but inflexible w.r.t. sensitive attributes

Further Motivation

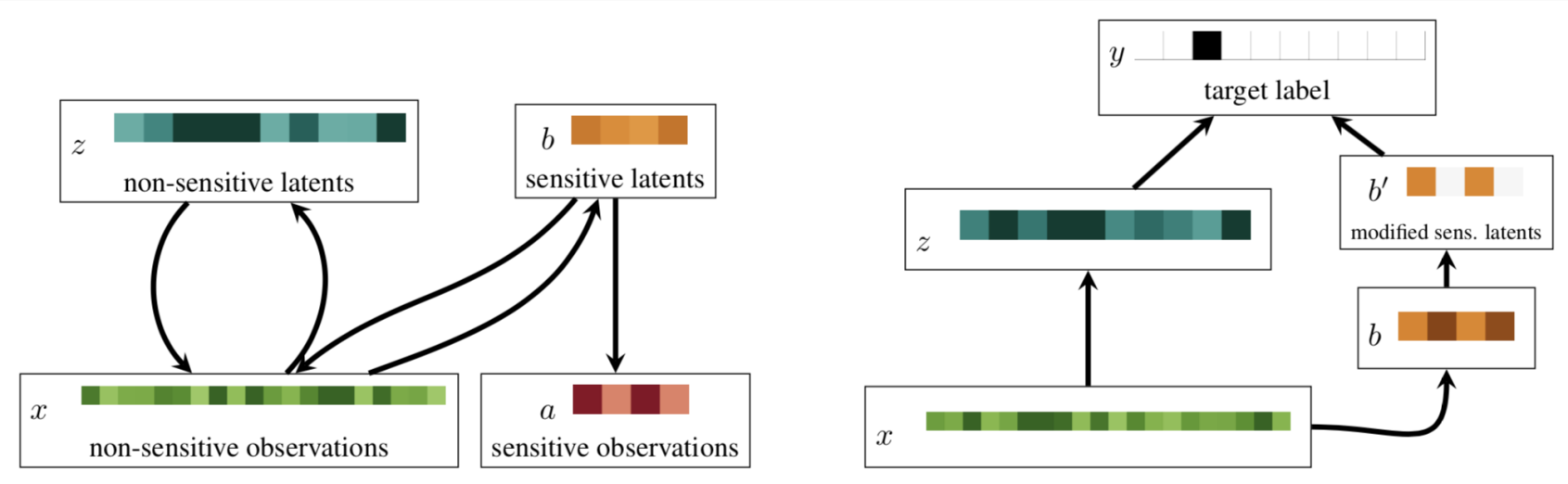
Subgroup discrimination

- We would like to handle the case where $\mathbf{a} \in \{0, 1\}^{N_a}$ is a vector of sensitive attributes
- ML systems can discriminate against subgroups defined via conjunctions of sensitive attributes (Buolamwini & Gebru, 2018)

Disentangled Representation Learning

- Each dimension of $\mathbf{z}$ should correspond to no more than one semantic factor of variation (object shape, position, etc.) in the data
- VAE variants encourage factorized posterior $q(\mathbf{z}|\mathbf{x})$ (Higgins et al., 2017) or aggregate posterior $q(\mathbf{z})$ (Kim & Mnih, 2018; Chen et al., 2018)

Flexibly Fair VAE



Data flow at train time (left) and test time (right) for FFVAE

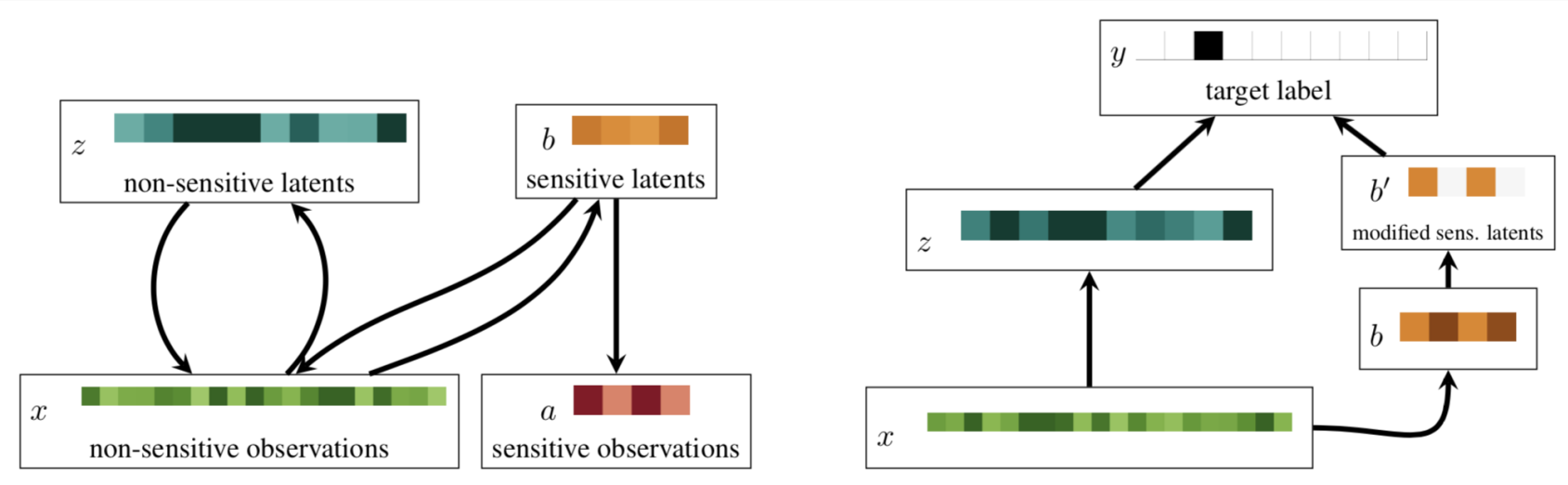
Desiderata

- $\mathbf{z} \perp b_j \; \forall \, j$
- $b_i \perp b_j \; \forall \, i \neq j$
- MutualInfo(a_j, b_j) is large $\forall \, j$

Latent Code Modification

- To achieve DP w.r.t. some a_i, use $[\mathbf{z}, \mathbf{b}] \backslash b_i$
- To achieve DP w.r.t. conjunction of binary attributes $a_i \wedge a_j \wedge a_k$, use $[\mathbf{z}, \mathbf{b}] \backslash \{b_i, b_j, b_k\}$

Flexibly Fair VAE



Data flow at train time (left) and test time (right) for FFVAE

Learning Objective

$$L_{\text{FFVAE}}(p,q) = \mathbb{E}_{q(\mathbf{z},\mathbf{b}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z},\mathbf{b}) + \alpha \sum_j \log p(a_j|b_j)]$$
$$- \gamma D_{KL}(q(\mathbf{z},\mathbf{b})||q(\mathbf{z})\prod_j q(b_j))$$
$$- D_{KL}\left[q(\mathbf{z},\mathbf{b}|\mathbf{x})||p(\mathbf{z},\mathbf{b})\right]$$

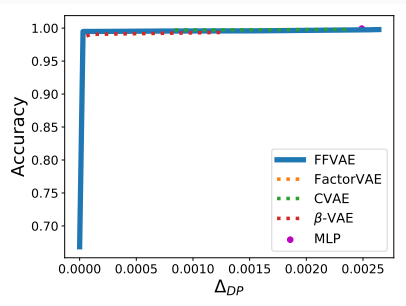
α encourages predictiveness in the latent code; γ encourages disentanglement

Results - Synthetic Data

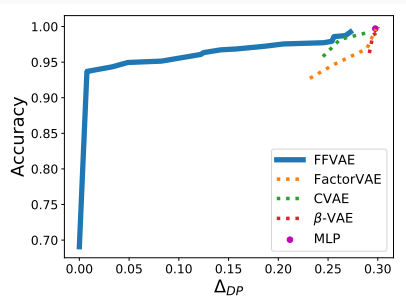
DSpritesUnfair

Samples 

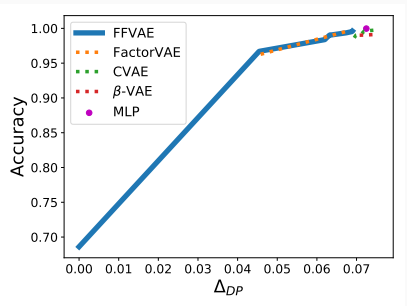
Figure: With correlated factors of variation, a fair classification task is predicting Shape without discriminating against XPosition



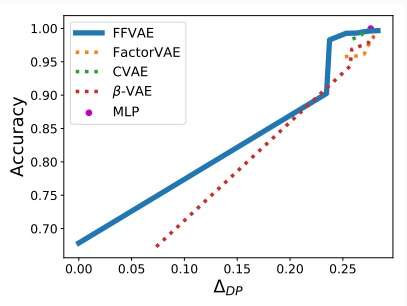
a = Scale



a = Shape



a=Shape$\wedge$Scale



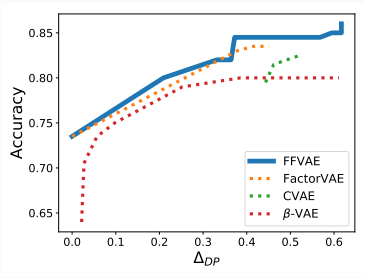
a=Shape$\vee$Scale

Figure: Pareto-fronts showing fairness-accuracy tradeoff curves, DSpritesUnfair dataset. Optimal point is top left corner (perfect accuracy, no unfairness). y = XPosition.

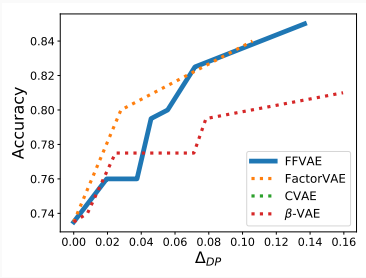
$$\Delta_{DP} \triangleq |\mathbb{E}[\hat{y} = 1|a = 1] - \mathbb{E}[\hat{y} = 1|a = 0]| \text{ with } \hat{y} \in \{0, 1\}$$

Results - Tabular and Image Data

Communities & Crime



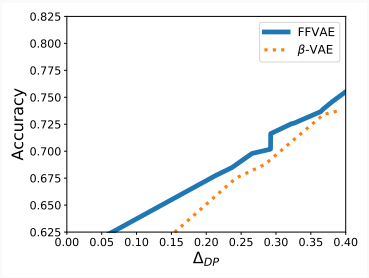
typical success:
$a = \mathrm{R} \wedge \mathrm{P}$



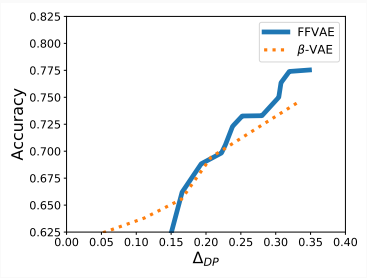
typical failure:
$a = \mathrm{R} \vee \mathrm{B}$

- Neighborhood-level population statistics: 120 features for $1,994$ neighborhoods
- We choose racePctBlack (R), blackPerCap (B), and pctNotSpeakEnglWell (P) as sensitive attributes
- Held-out label
  $y =$ violentCrimesPerCapita

Celeb-A



typical success:
$a = \neg\, \mathrm{E} \wedge \mathrm{M}$



typical failure:
$a = \mathrm{C} \wedge \mathrm{M}$

- Over $200,000$ images of celebrity faces, each associated with 40 binary attributes (OvalFace, HeavyMakeup, etc.)
- We choose Chubby (C), Eyeglasses (E) and Male (M) as sensitive attributes
- Held-out label
  $y =$ HeavyMakeup (H)

Conclusion

- FFVAE enables flexibly fair downstream classification by disentangling information from multiple sensitive attributes
- Future work: extending to other group fairness definitions, and studying robustness of disentangled/fair representation learners to distribution shift
- Visit us at **poster # 131** tonight!