Disentangled Fair Representations

May 29, 2020

Dr. Wei Wei, Prof. James Landay

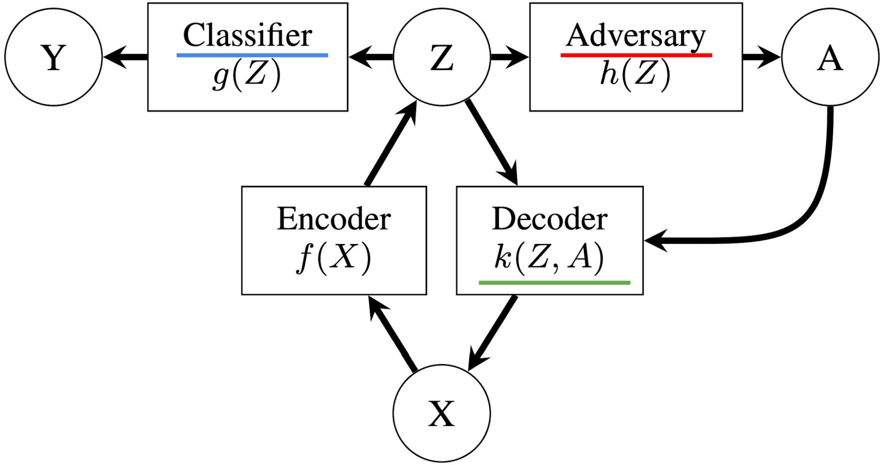
CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Stanford University

Fairness Through Adversarial Learning

- Adversarial Learning

$$L(f,g,h,k) = \alpha L_C(g(f(X,A)),Y) + \beta L_{Dec}(k(f(X,A),A),X) + \gamma L_{Adv}(h(f(X,A)),A)$$

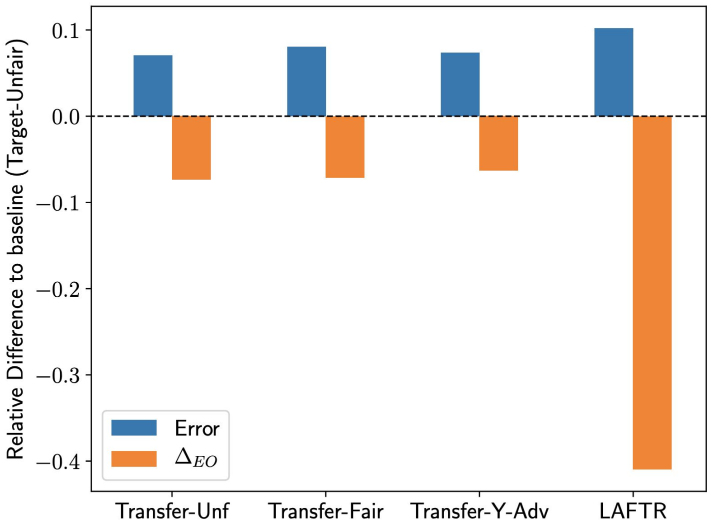


$$\underset{f,g,k}{\text{minimize}}\ \underset{h}{\text{maximize}}\ \mathbb{E}_{X,Y,A}\left[L(f,g,h,k)\right]$$

Madras et al, 2018

Transfer Fair Representations

- Heritage Health Dataset
  - Comprises insurance claims and physician records
  - Task 1 - Predict Charlson index (prediction of 10 year survival of patients) trained using equalized odds adversarial objective
  - Task 2 - Same input, task becomes predicting a patient’s insurance claim corresponding to a specific medical condition



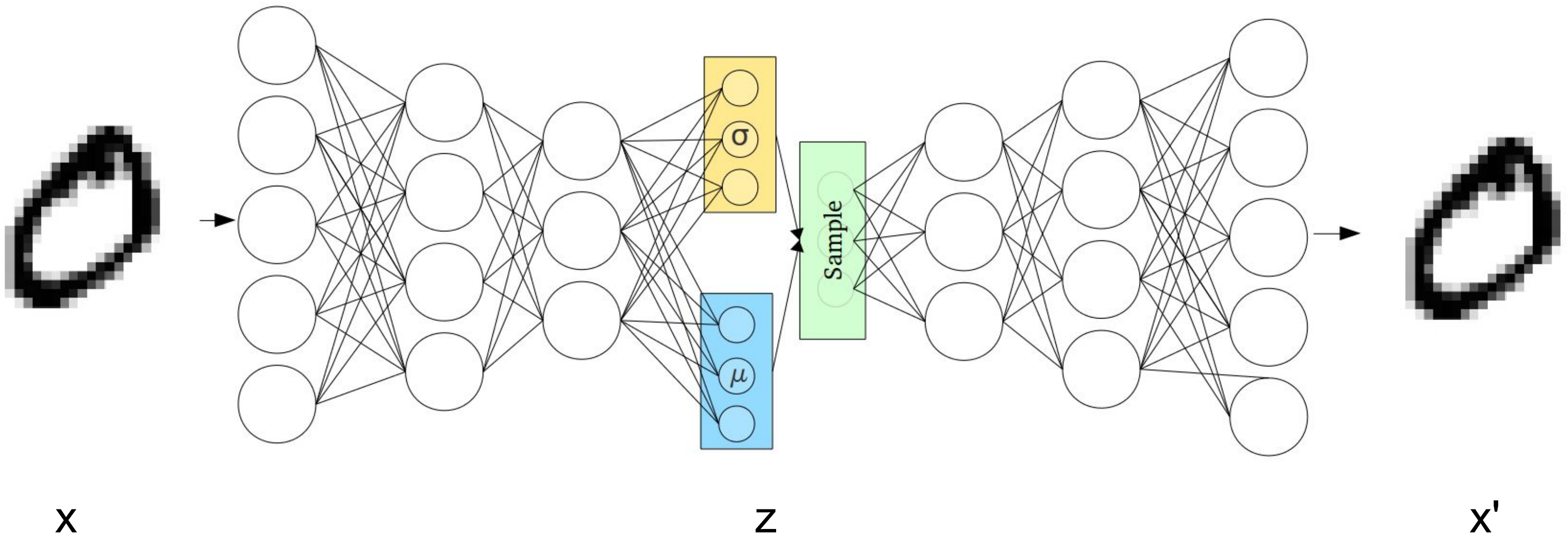
Transfer- unf - MLP with no fairness constraints
Transfer- fair - MLP with fairness constraints in Bechavod et al, 2017
Transfer - Y - Adv baseline in Zhang et al, 2018

Madras et al, 2018

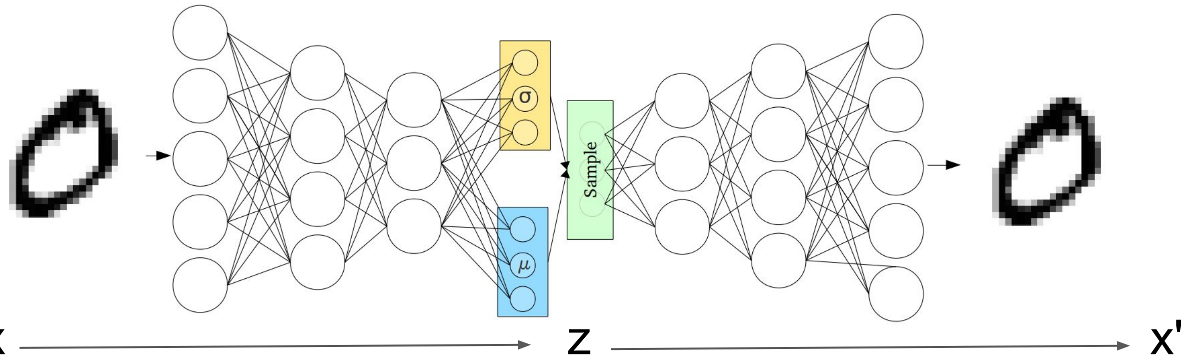
Outline

- Disentangled Representations
- Flexibly Fair Representation
- Orthogonal Disentangled Fair Representations
- Measurements for Disentangled Fair Representations

VAE Revisited



VAE Revisited



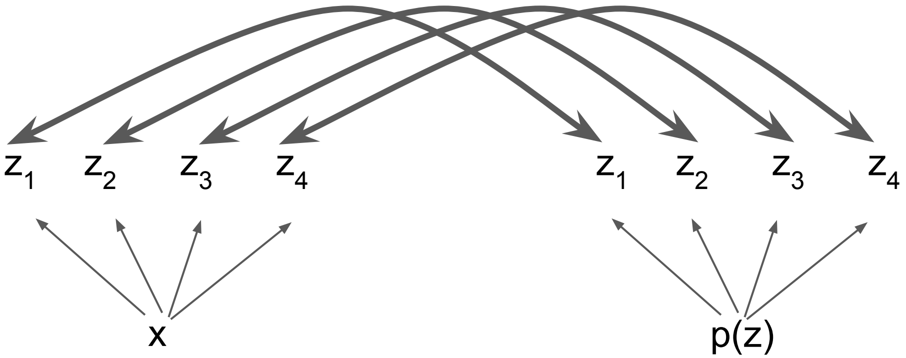
$$q(z|x) = \mathcal{N}(z|\mu_q(x), \Sigma_q(x)) \qquad p(x|z) = \mathcal{N}(x|\mu_p(z), \Sigma_p(z))$$

$$p(z) = \mathcal{N}(0, I)$$

$$L_{\text{VAE}}(p, q) = \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] - D_{KL}\left[q(z|x)||p(z)\right]$$

Disentanglement in VAE

$$L_{\text{VAE}}(p,q) = \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] - D_{KL}\left[q(z|x)||p(z)\right]$$



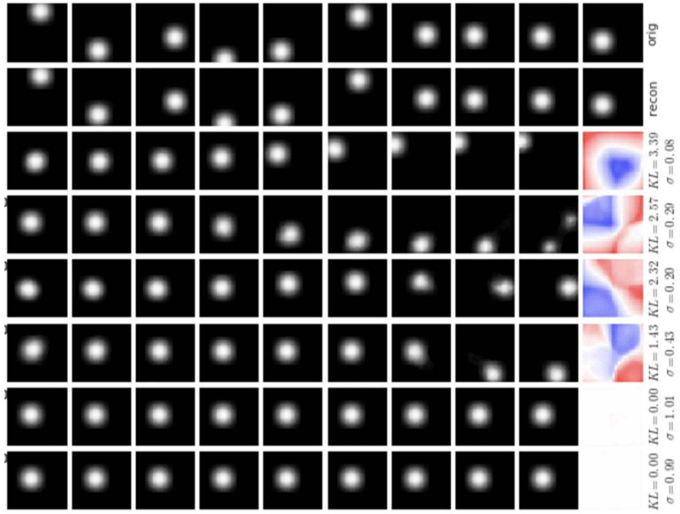
$$p(z) = \prod_j p(z_j)$$

Higgins et al, 2017

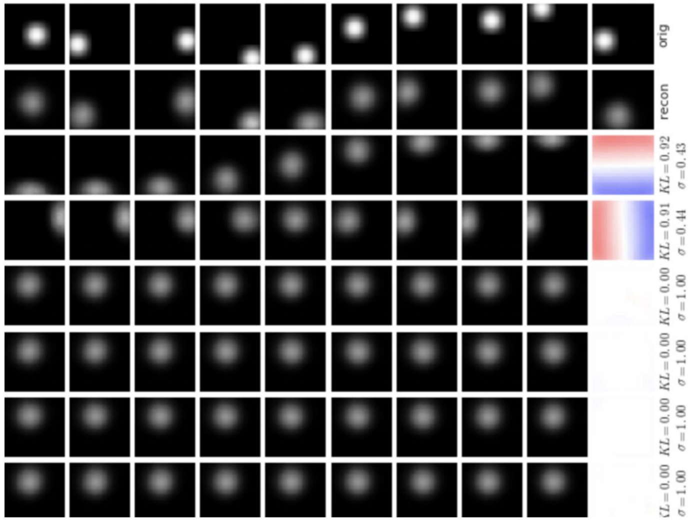
β-VAE

$$L_{\beta\text{VAE}}(p, q) = \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] - \beta D_{KL}\left[q(z|x)||p(z)\right]$$

β = 1



β = 150



[Matthey et al, 2018]

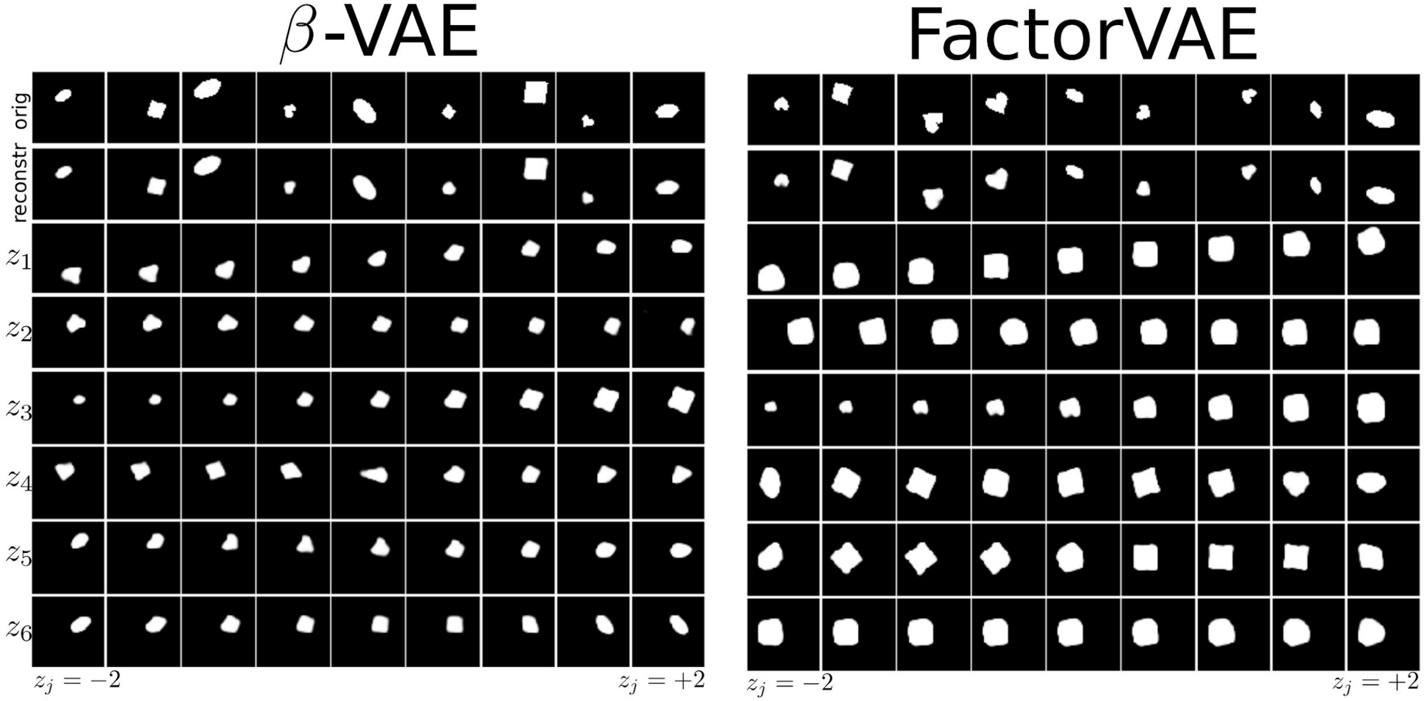
FactorVAE

$$L_{\text{FactorVAE}}(p,q) = \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] - D_{KL}\left[q(z|x)||p(z)\right] - \gamma D_{KL}(q(z)||\prod_j q(z_j))$$

z_i correlates with z_j if and only if i = j

Kim et al, 2018

β-VAE and FactorVAE

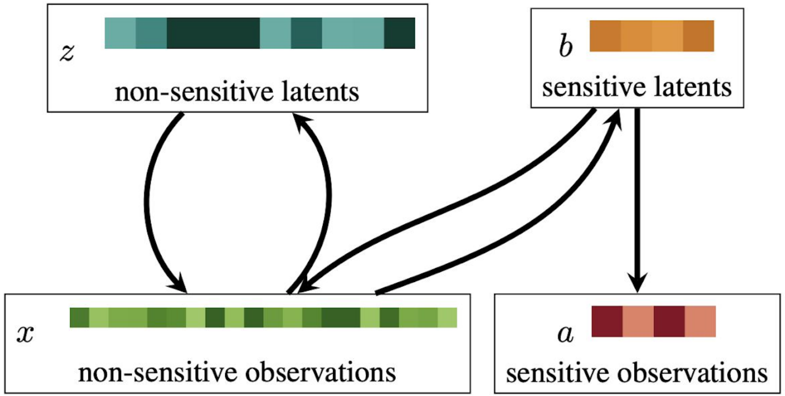


Models find x-position, y-position, and scale, but struggle to disentangle orientation and shape

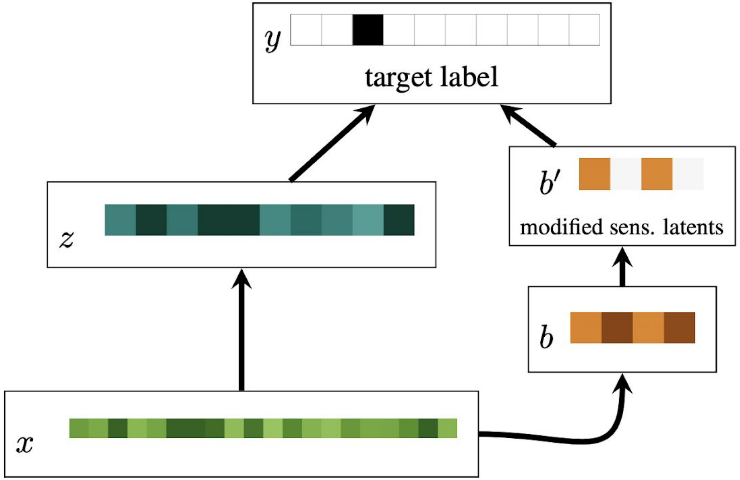
Outline

- Disentangled Representations
- Flexibly Fair Representation
- Orthogonal Disentangled Fair Representations
- Measurements for Disentangled Fair Representations

Flexibly Fair Representation



Training



Testing

Creager et al, 2019

Disentangled Fair Representations

$$q(z,b) = q(z)\prod_j q(b_j)$$

- Demographic Parity for Feature a_i
  - Ignoring a_i, use instead $[z, b]\backslash b_i$
  - or replace b_i with independent noise

- Compositional Procedure
  - use representation $[z, b]\backslash\{b_i , b_j , b_k\}$ for fair combination $\{a_i, a_j , a_k\}$

z - non-sensitive dimension of the latent variables
b - sensitive dimensions of the latent variables

Creager et al, 2019

Flexibly Fair Representation

- $z \perp b_j \; \forall \; j$ (disentanglement of the non-sensitive and sensitive latent dimensions);
- $b_i \perp b_j \; \forall \; i \neq j$ (disentanglement of the various different sensitive dimensions);
- $\mathrm{MI}(a_j, b_j)$ is large $\forall \, j$ (predictiveness of each sensitive dimension);

Creager et al, 2019

Flexibly Fair Representation

$$L_{\text{FFVAE}}(p,q) = \mathbb{E}_{q(\underline{z,b}|x,a)}\left[\log p(x,a|z,b)\right]$$

$z \perp b_j$

$p(z,b) = p(z)p(b)$

Standard Gaussian Uniform

$$\underbrace{- D_{KL}\left[q(z,b|x)||p(z,b)\right]}_{\beta\text{-VAE}}$$

β-VAE

$b_i \perp b_j \, \forall \, i \neq j$

$$\underbrace{- \gamma D_{KL}(q(z,b)||q(z)\prod_j q(b_j))}_{\text{factor-VAE}}$$

factor-VAE

[Creager et al, 2019]

Flexibly Fair Representation

$$\mathbb{E}_{q(z,b|x,a)}\left[\log p(x,a|z,b)\right] \Longrightarrow \mathbb{E}_{q(z,b|x)}\left[\log p(x|z,b) + \alpha \log p(a|b)\right]$$

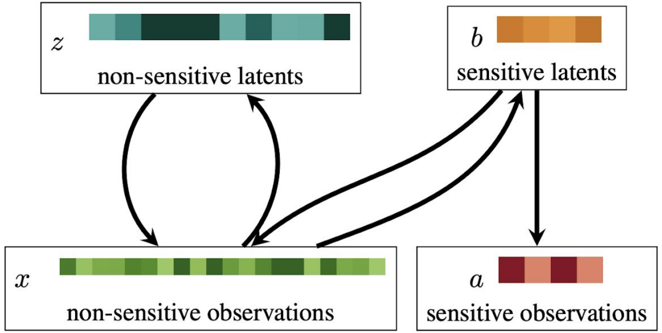
$$p(x,a|z,b) = p(x|z,b)p(a|b)$$

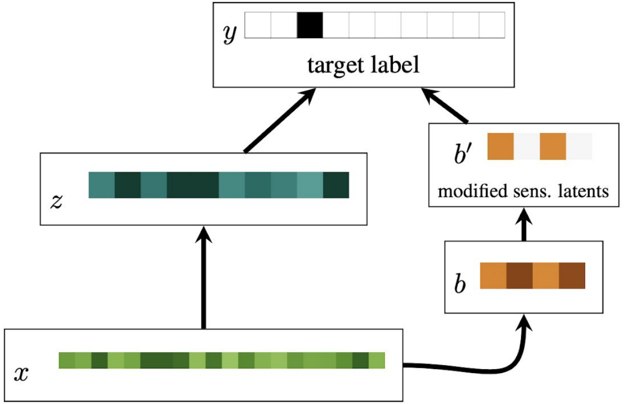
$$\begin{aligned} L_{\text{FFVAE}}(p,q) = \mathbb{E}_{q(z,b|x)}&\left[\log p(x|z,b) + \alpha \log p(a|b)\right] \\ &- \gamma D_{KL}\Big(q(z,b)||q(z)\prod_j q(b_j)\Big) \\ &- D_{KL}\left[q(z,b|x)||p(z,b)\right]. \end{aligned}$$

Creager et al, 2019

Experiments

- Fair Classification
  - Make fair predictions
- Predictiveness
  - Train a classifier to predict sensitive attribute a_i from b_i alone
- Disentanglement
  - Train a classifier to predict sensitive attribute a_i from representations with b_i removed





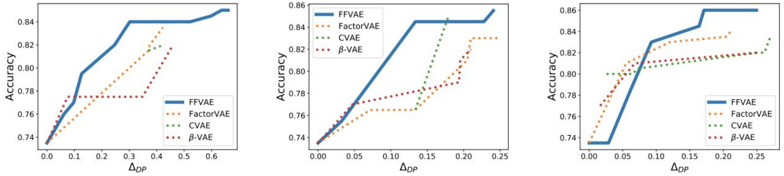
Communities & Crime

Fair Classification

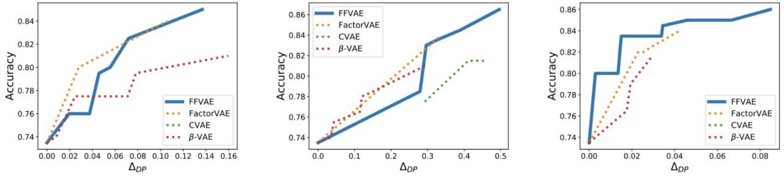
$$\Delta_{DP}(g) \triangleq d_g(\mathcal{Z}_0, \mathcal{Z}_1) = |\mathbb{E}_{\mathcal{Z}_0}[g] - \mathbb{E}_{\mathcal{Z}_1}[g]|$$

$$\Delta_{DP}(g) = 0 \iff g(Z) \perp\!\!\!\perp A$$

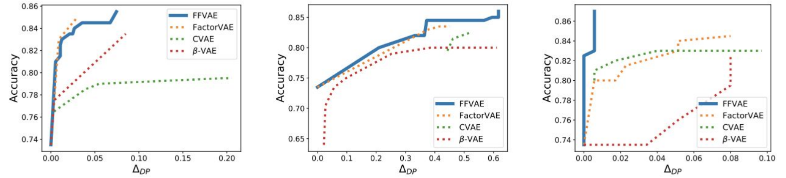
- Sensitive attributes:
    - racePctBlack (R)
    - blackPerCapIncome (B)
    - pctNotSpeakEnglWell (P)
- y = violentCrimesPerCaptia



(a) $a = \mathrm{R}$ (b) $a = \mathrm{B}$ (c) $a = \mathrm{P}$



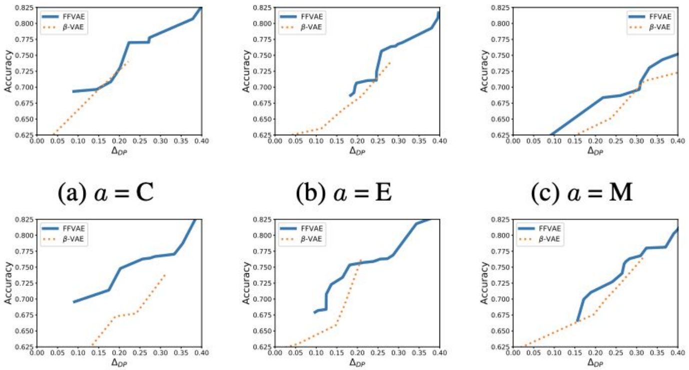
(d) $a = \mathrm{R} \vee \mathrm{B}$ (e) $a = \mathrm{R} \vee \mathrm{P}$ (f) $a = \mathrm{B} \vee \mathrm{P}$



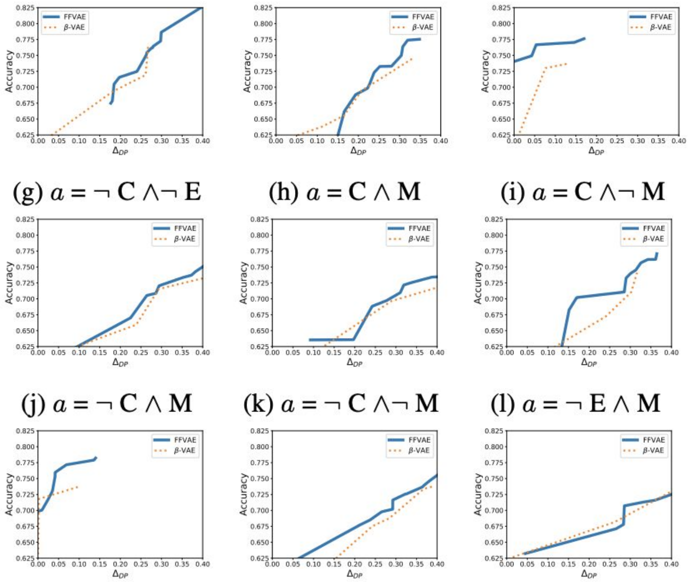
(g) $a = \mathrm{R} \wedge \mathrm{B}$ (h) $a = \mathrm{R} \wedge \mathrm{P}$ (i) $a = \mathrm{B} \wedge \mathrm{P}$

CelebA

Fair Classification



(a) $a = \text{C}$ (b) $a = \text{E}$ (c) $a = \text{M}$



(d) $a = \text{C} \wedge \text{E}$ (e) $a = \text{C} \wedge \neg\, \text{E}$ (f) $a = \neg\, \text{C} \wedge \text{E}$

(g) $a = \neg\, \text{C} \wedge \neg\, \text{E}$ (h) $a = \text{C} \wedge \text{M}$ (i) $a = \text{C} \wedge \neg\, \text{M}$

(j) $a = \neg\, \text{C} \wedge \text{M}$ (k) $a = \neg\, \text{C} \wedge \neg\, \text{M}$ (l) $a = \neg\, \text{E} \wedge \text{M}$

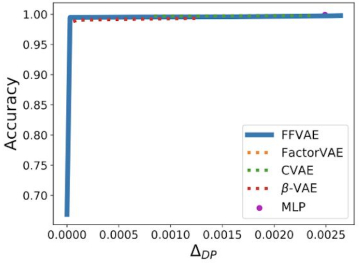
(m) $a = \text{E} \wedge \neg\, \text{M}$ (n) $a = \neg\, \text{E} \wedge \text{M}$ (o) $a = \neg\, \text{E} \wedge \neg\, \text{M}$

- Sensitive attributes
  - Chubby (C)
  - Eyeglasses (E)
  - Male (M)
- y = HeavyMakeup.

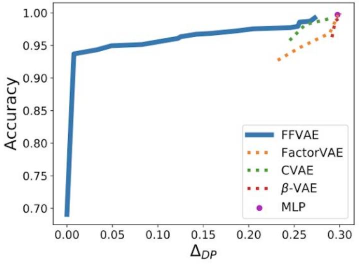
DSpritesUnfair Dataset

Fair Classification

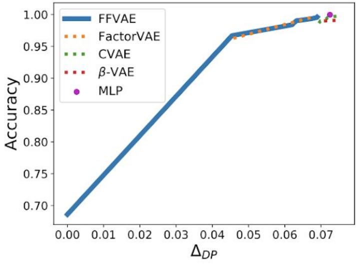
2D shapes procedurally generated from 6 ground truth independent latent factors. These factors are color, shape, scale, rotation, x and y positions of a sprite.



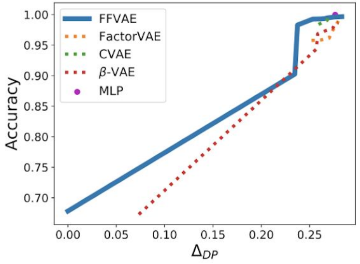
(a) a = Scale



(b) a = Shape



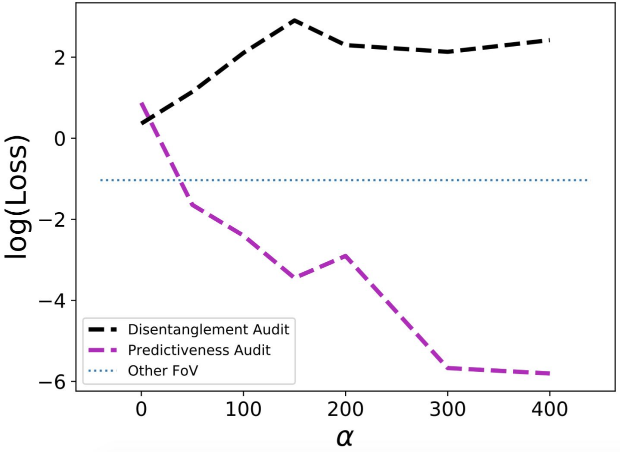
(c) a = Shape $\wedge$ Scale



(d) a = Shape $\vee$ Scale

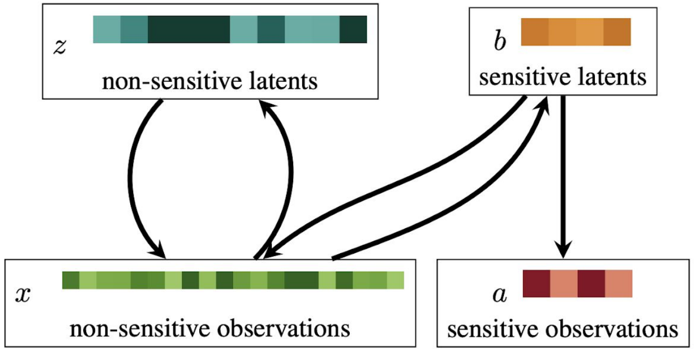
DSpritesUnfair Dataset

- Disentanglement - Predict sensitive attribute a_i from b_i alone
- Predictiveness - Predict sensitive attribute a_i from representations with b_i removed

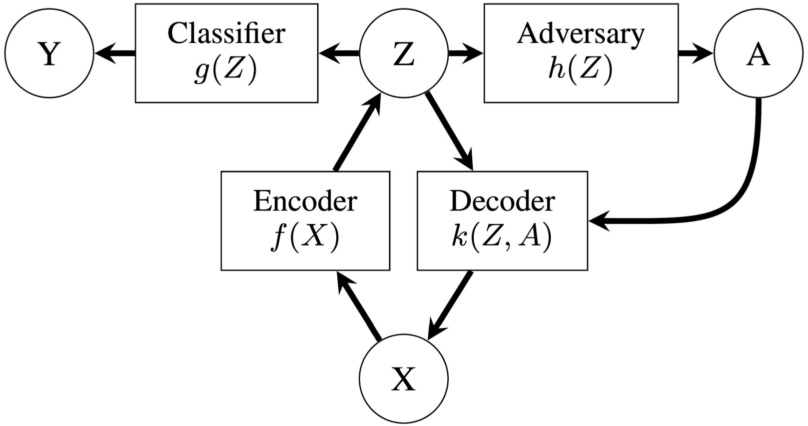


$$L_{\text{FFVAE}}(p,q) = \mathbb{E}_{q(z,b|x)}[\log p(x|z,b) + \underline{\alpha \log p(a|b)}] \\ - \gamma D_{KL}(q(z,b)||q(z)\prod_j q(b_j)) \\ - D_{KL}\left[q(z,b|x)||p(z,b)\right].$$

Comparisons to Adversarial Learning



Flexibly Fair Representation



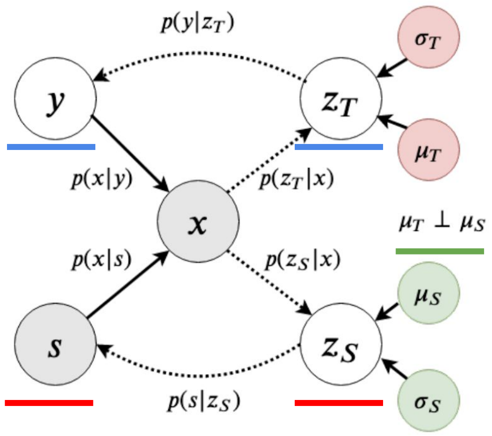
Adversarial Learning

Outline

- Disentangled Representations
- Flexibly Fair Representation
- Orthogonal Disentangled Fair Representations
- Measurements for Disentangled Fair Representations

Orthogonal Disentangled Fair Representations

- Train a fair representation that is
  - Disentangled
  - and Orthogonal



Sarhan et al, 2020

Training Objective

$$arg \min_{\theta_T, \theta_S, \phi_T, \phi_S} \underline{\mathcal{L}_T(\theta_T, \phi_T)} + \underline{\mathcal{L}_S(\theta_S^*, \phi_S)} + \lambda_E \mathcal{L}_E(\phi_S, \theta_T) + \lambda_{OD} \mathcal{L}_{OD}(\theta_T, \theta_S)$$

$$\mathcal{L}_T(\theta_T, \phi_T) = \mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x}) \; || \; q_{\phi_T}(\boldsymbol{y}|\boldsymbol{z_T}))$$
$$\mathcal{L}_S(\theta_S^*, \phi_S) = \mathrm{KL}(p(\boldsymbol{s}|\boldsymbol{x}) \; || \; q_{\phi_S}(\boldsymbol{s}|\boldsymbol{z_S}))$$

Matching the probabilities of ground truth (i.e., y) and sensitive information (i.e., s).

Training Objective

$$arg \min_{\theta_T, \theta_S, \phi_T, \phi_S} \mathcal{L}_T(\theta_T, \phi_T) + \mathcal{L}_S(\theta_S^*, \phi_S) + \underline{\lambda_E \mathcal{L}_E(\phi_S, \theta_T)} + \lambda_{OD} \mathcal{L}_{OD}(\theta_T, \theta_S)$$

$$\mathcal{L}_E(\phi_S, \theta_T) = \mathrm{KL}(q_{\phi_S}(\boldsymbol{s}|\boldsymbol{z_T}) \;||\; \mathcal{U}(\boldsymbol{s}))$$

Makes sure that none sensitive information got leaked into the prediction z_T

Training Objective

$$arg \min_{\theta_T, \theta_S, \phi_T, \phi_S} \mathcal{L}_T(\theta_T, \phi_T) + \mathcal{L}_S(\theta_S^*, \phi_S) + \lambda_E \mathcal{L}_E(\phi_S, \theta_T) + \underline{\lambda_{OD} \mathcal{L}_{OD}(\theta_T, \theta_S)}$$

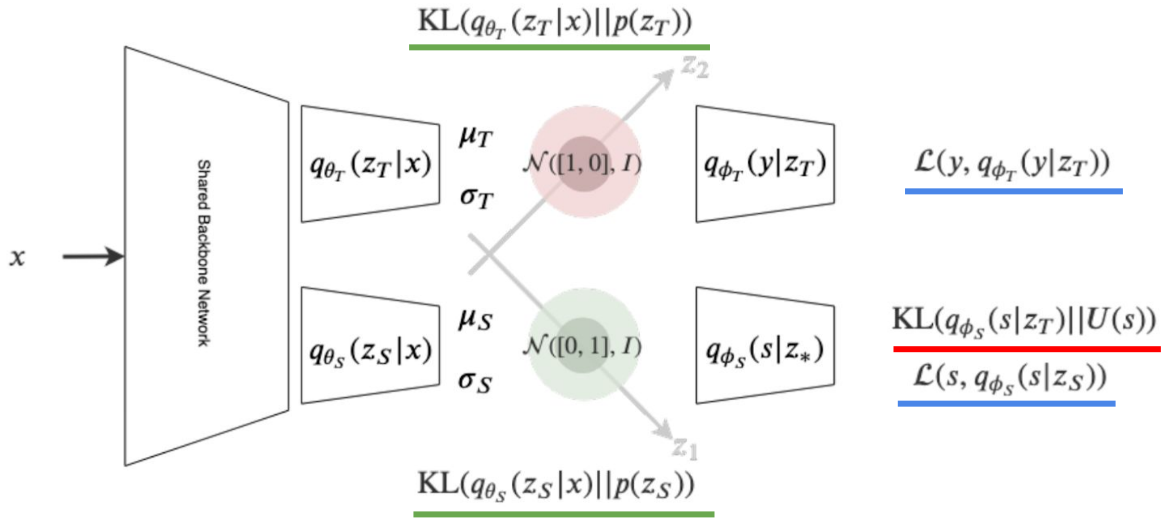
$$\mathcal{L}_{OD}(\theta_T, \theta_S) = \mathcal{L}_{\boldsymbol{z_T}}(\theta_T) + \mathcal{L}_{\boldsymbol{z_S}}(\theta_S)$$

$$\mathcal{L}_{\boldsymbol{z_T}}(\theta_T) = \mathrm{KL}(q_{\theta_T}(\boldsymbol{z_T}|\boldsymbol{x}) \mid\mid p(\boldsymbol{z_T}))$$

$$p(\boldsymbol{z_S}) = \mathcal{N}([0,1]^T, I) \quad p(\boldsymbol{z_T}) = \mathcal{N}([1,0]^T, I)$$

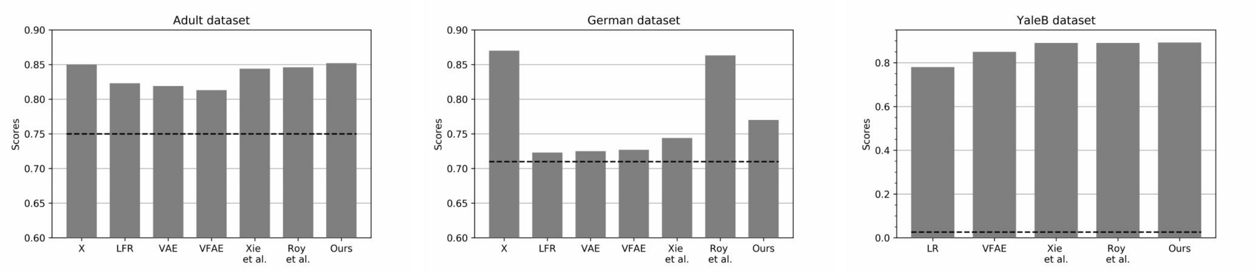
Enforces both Disentanglement and Orthogonality

Training Objective

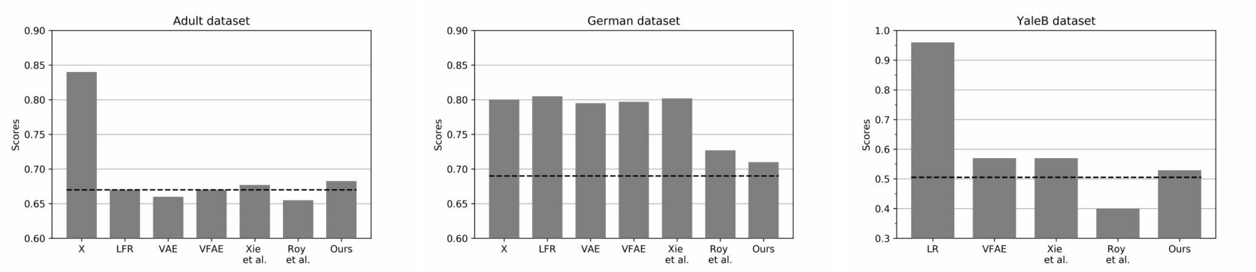


$$arg \min_{\theta_T,\theta_S,\phi_T,\phi_S} \mathcal{L}_T(\theta_T,\phi_T) + \mathcal{L}_S(\theta_S^*,\phi_S) + \lambda_E\mathcal{L}_E(\phi_S,\theta_T) + \lambda_{OD}\mathcal{L}_{OD}(\theta_T,\theta_S)$$

Adult, German, and extended YaleB



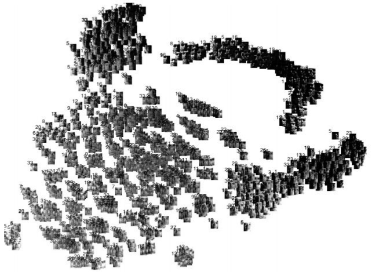
(a) Target attribute classification accuracy.



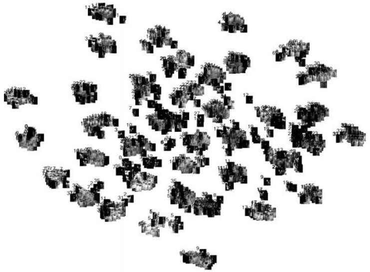
(b) Sensitive attribute classification accuracy.

Visualizations on the Embeddings

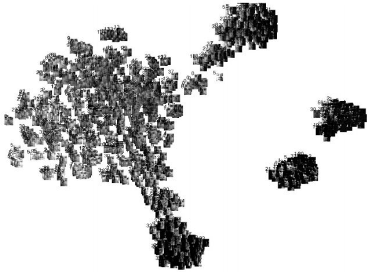
YaleB faces



(a) t-SNE on x



(b) t-SNE on $\boldsymbol{z_T}$



(c) t-SNE on $\boldsymbol{z_S}$

CIFAR 10

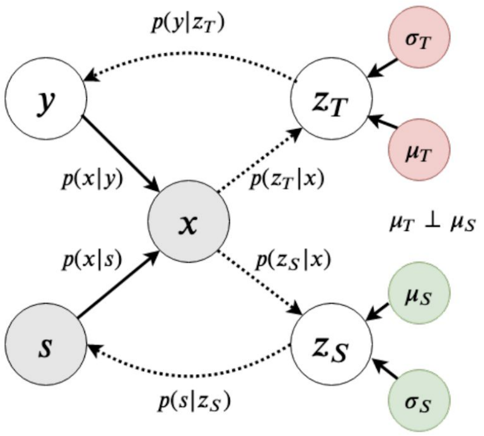
(d) t-SNE on x

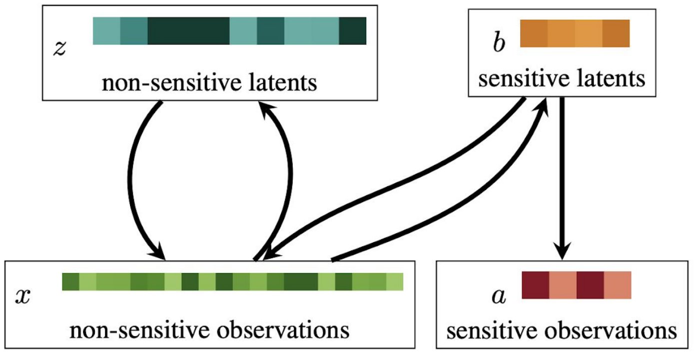
(e) t-SNE on $\boldsymbol{z_T}$

(f) t-SNE on $\boldsymbol{z_S}$

Comparisons to Flexibly Fair Representation

- How do they handle leakage of sensitive information to the representations?
- How do they handle disentanglement?

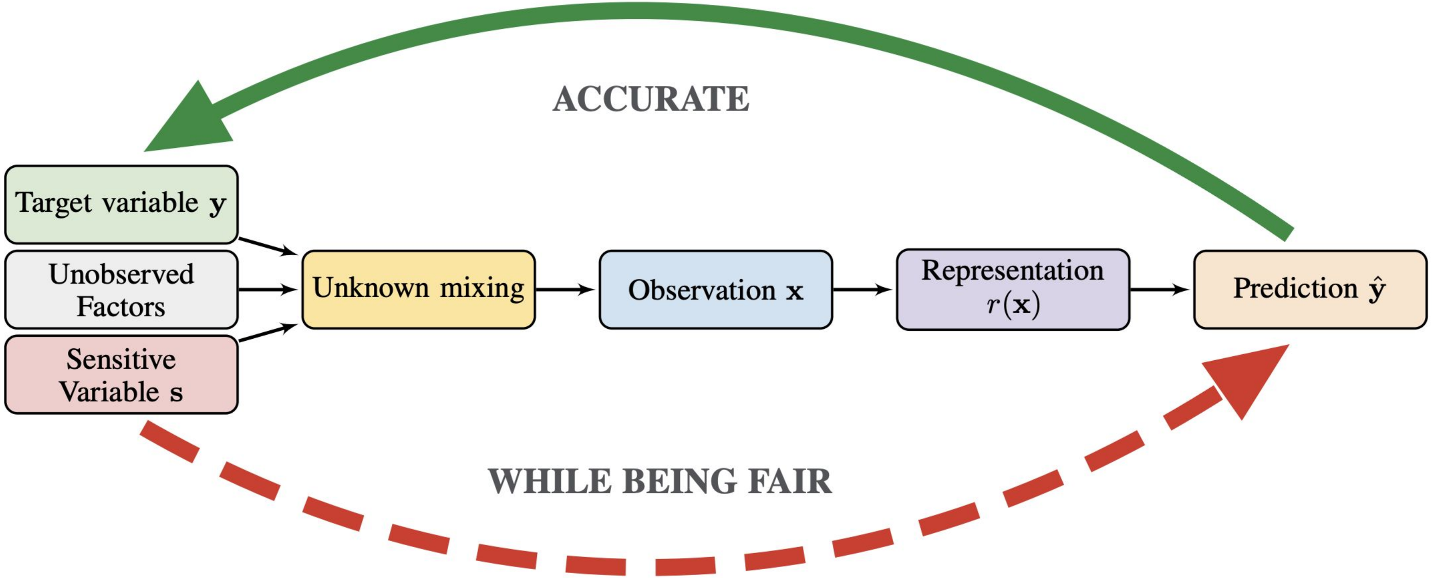




Outline

- Disentangled Representations
- Flexibly Fair Representation
- Orthogonal Disentangled Fair Representations
- Measurements for Disentangled Fair Representations

Measurements for Disentangled Fair Representations



[Locatello et al, 2019]()

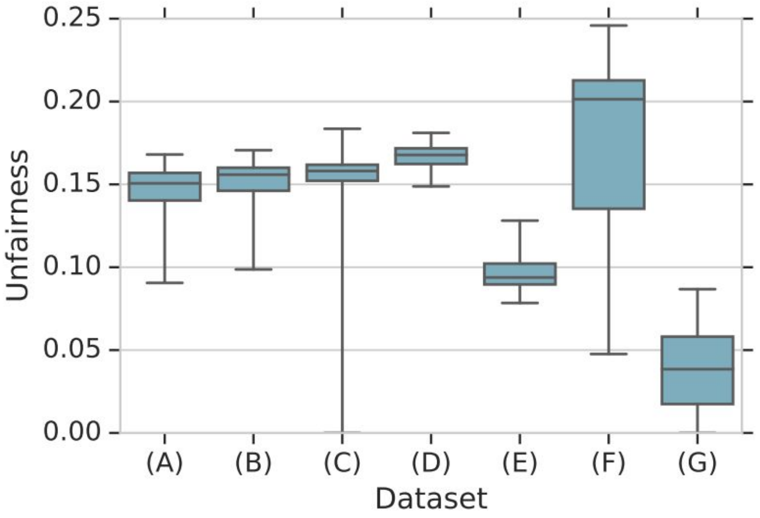
Unfairness Measure

- Measuring Unfairness Without Ground Truth
  - Total Variation (TV) of prediction pairs across groups

$$\texttt{unfairness}(\hat{\mathbf{y}}) = \frac{1}{|S|} \sum_{s} TV\left(p(\hat{\mathbf{y}}), p(\hat{\mathbf{y}} \mid \mathbf{s} = s)\right) \; \forall \, y$$

Locatello et al, 2019

Results Using Models Trained in [Locatello et al, 2019]



	dSprites	Color-dSprites	Noisy-dSprites	Scream-dSprites	SmallNORB	Cars3D	Shapes3D
BetaVAE Score	-69	-72	-20	-64	-57	-15	-50
FactorVAE Score	-69	-68	-41	-56	-55	-19	-40
MIG	-86	-81	-47	-65	-48	-32	-71
DCI Disentanglement	-92	-86	-71	-77	-55	-55	-95
Modularity	4	-6	28	4	41	-18	-55
SAP	-70	-64	-15	-60	-9	6	-53

A - dSprites, B - Color-dSprites, C - Noisy-dSprites
D - Scream-dSprites , E - SmallNORB, F - Cars3D, G - Shapes3D

Locatello et al, 2019

Outline

- Disentangled Representations
- Flexibly Fair Representation
- Orthogonal Disentangled Fair Representations
- Measurements for Disentangled Fair Representations

Reading Assignments (Disentangled Fair Representations)

- Zhao, Han, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations, ICLR 2020
- Zhao, Han, and Geoff Gordon. Inherent tradeoffs in learning fair representations, NeurIPS 2019
- He, Yuzi, Keith Burghardt, and Kristina Lerman. A Geometric Solution to Fair Representations, AAAI/ACM AI, Ethics, and Society 2020
- Ruoss, Anian, Mislav Balunović, Marc Fischer, and Martin Vechev. Learning Certified Individually Fair Representations, arXiv 2020
- Chiappa, Silvia, Ray Jiang, Tom Stepleton, Aldo Pacchiano, Heinrich Jiang, and John Aslanides. A general approach to fairness with optimal transport, AAAI 2020