Fairness Through Adversarial Learning

- Adversarial Learning

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

\[
\begin{align*}
Y &\xleftarrow{g(Z)} Z & \xrightarrow{h(Z)} A \\
X & \quad f(X) & \quad k(Z, A) \\
& \quad \text{Encoder} & \quad \text{Decoder} \\
& \quad \text{Classifier} & \quad \text{Adversary} \\
& \quad \text{minimize} & \quad \text{maximize} \\
& \quad \mathbb{E}_{X,Y,A} [L(f, g, h, k)] & \quad \mathbb{E}_h \end{align*}
\]

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

- ML Auditing and Accountability
  - Distill-and-Compare
VAE Revisited
VAE Revisited

\[
q(z|x) = \mathcal{N}(z|\mu_q(x), \Sigma_q(x)) \quad p(x|z) = \mathcal{N}(x|\mu_p(z), \Sigma_p(z)) \quad p(z) = \mathcal{N}(0, I)
\]

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

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

\[ p(z) = \prod_j p(z_j) \]

Higgins et al, 2017
$L_{\beta VAE}(p, q) = \mathbb{E}_{q(z|x)} [\log p(x|z)] - \beta D_{KL} [q(z|x) \| p(z)]$

$\beta = 1$

$\beta = 150$
FactorVAE

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

\[ z_i \text{ correlates with } z_j \text{ if and only if } i = j \]

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 Fair Representations
  - Disentangled Representations
  - Flexibly Fair Representation
  - Orthogonal Disentangled Fair Representations
  - Measurements for Disentangled Fair Representations
- ML Auditing
  - Distill-and-Compare
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 \([z, b] \setminus b_i\)
  - or replace \( b_i \) with independent noise

- Compositional Procedure
  - use representation \([z, b] \setminus \{b_i, b_j, b_k\}\) 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);
- $\text{MI}(a_j, b_j)$ is large $\forall \; j$ (predictiveness of each sensitive dimension);
Flexibly Fair Representation

\[
L_{FFVAE}(p, q) = \mathbb{E}_{q(z,b|x,a)} \left[ \log p(x, a | z, b) \right] - D_{KL} [ q(z, b|x) || p(z, b) ] - \gamma D_{KL} (q(z, b) || q(z) \prod_{j} q(b_{j}))
\]

- \( z \perp b_{j} \)
- \( p(z, b) = p(z)p(b) \)  \text{Standard Uniform Gaussian}
- \( b_{i} \perp b_{j} \forall i \neq j \)

\text{Creager et al, 2019}
Flexibly Fair Representation

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

\[ p(x,a|z,b) = p(x,z,b)p(a|b) \]

\[ L_{\text{FFVAE}}(p,q) = \mathbb{E}_{q(z,b|x)} [\log p(x|z, b) + \alpha \log p(a|b)] \]

\[ - \gamma D_{KL}(q(z,b)||q(z) \prod_j q(b_j)) \]

\[ - D_{KL} [q(z,b|x)||p(z,b)] . \]

*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

$\Delta_{DP}(g) \triangleq d_g(Z_0, Z_1) = \left| \mathbb{E}_{Z_0}[g] - \mathbb{E}_{Z_1}[g] \right|$  

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

- Sensitive attributes:
  - racePctBlack (R)
  - blackPerCapIncome (B)
  - pctNotSpeakEnglWell (P)
- $y = \text{violentCrimesPerCaptia}$
CelebA

Fair Classification

- Sensitive attributes
  - Chubby (C)
  - Eyeglasses (E)
  - Male (M)
- $y = \text{HeavyMakeup}$.
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.
DSpritesUnfair Dataset

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

$$L_{FFVAE}(p, q) = \mathbb{E}_{q(z,b|x)}[\log p(x|z, b) + \alpha \log p(a|b)]$$

$$- \gamma D_{KL}(q(z, b)\|q(z) \prod_j q(b_j))$$

$$- D_{KL}[q(z, b|x)\|p(z, b)].$$
Comparisons to Adversarial Learning

Flexibly Fair Representation

Adversarial Learning
Outline

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

- ML Auditing
  - Distill-and-Compare
Orthogonal Disentangled Fair Representations

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

Sarhan et al, 2020
Training Objective

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

$$\mathcal{L}_T(\theta_T, \phi_T) = \text{KL}(p(y|x) \parallel q_{\phi_T}(y|z_T))$$

$$\mathcal{L}_S(\theta_S^*, \phi_S) = \text{KL}(p(s|x) \parallel q_{\phi_S}(s|z_S))$$

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

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

\[
\mathcal{L}_E(\phi_S, \theta_T) = KL(q_{\phi_S}(s|z_T) \parallel \mathcal{U}(s))
\]

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

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

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

$$\mathcal{L}_{z_T}(\theta_T) = \text{KL}(q_{\theta_T}(z_T|x) \parallel p(z_T))$$

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

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 $z_T$
(c) t-SNE on $z_S$

CIFAR 10
(d) t-SNE on $x$
(e) t-SNE on $z_T$
(f) t-SNE on $z_S$
Comparisons to Flexibly Fair Representation

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

● Disentangled Fair Representations
  ○ Disentangled Representations
  ○ Flexibly Fair Representation
  ○ Orthogonal Disentangled Fair Representations
  ○ Measurements for Disentangled Fair Representations

● ML Auditing
  ○ Distill-and-Compare
Measurements for Disentangled Fair Representations

Locatello et al, 2019
Unfairness Measure

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

\[
\text{unfairness}(\hat{y}) = \frac{1}{|S|} \sum_s TV(p(\hat{y}), p(\hat{y} \mid s = s)) \forall y
\]

Locatello et al, 2019
Results Using Models Trained in **Locatello et al, 2019**

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BetaVAE Score</th>
<th>FactorVAE Score</th>
<th>MIG</th>
<th>DCI Disentanglement</th>
<th>Modularity</th>
<th>SAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>dSprites</td>
<td>-69</td>
<td>-69</td>
<td>-86</td>
<td>-92</td>
<td>4</td>
<td>-70</td>
</tr>
<tr>
<td>Color-dSprites</td>
<td>-72</td>
<td>-68</td>
<td>-81</td>
<td>-86</td>
<td>-6</td>
<td>-64</td>
</tr>
<tr>
<td>Noisy-dSprites</td>
<td>-20</td>
<td>-41</td>
<td>-47</td>
<td>-71</td>
<td>28</td>
<td>-15</td>
</tr>
<tr>
<td>Scream-dSprites</td>
<td>-64</td>
<td>-56</td>
<td>-65</td>
<td>-77</td>
<td>4</td>
<td>-60</td>
</tr>
<tr>
<td>SmallNORB</td>
<td>-57</td>
<td>-55</td>
<td>-65</td>
<td>-55</td>
<td>41</td>
<td>-9</td>
</tr>
<tr>
<td>Shapes3D</td>
<td>-50</td>
<td>-40</td>
<td>-71</td>
<td>-95</td>
<td>-55</td>
<td>-53</td>
</tr>
</tbody>
</table>

**Locatello et al, 2019**
Outline

● Disentangled Fair Representations
  ○ Disentangled Representations
  ○ Flexibly Fair Representation
  ○ Orthogonal Disentangled Fair Representations
  ○ Measurements for Disentangled Fair Representations

● ML Auditing
  ○ Distill-and-Compare
Accountability

- Who takes the responsibilities for failed ML models?
Regulating AI Models

<table>
<thead>
<tr>
<th>Attribute</th>
<th>FHA</th>
<th>ECOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Color</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>National origin</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Religion</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Sex</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Familial status</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Disability</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Exercised rights under CCPA</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Marital status</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Recipient of public assistance</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Age</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
ML Auditing Using Model Distillation

Tan et al., 2018
General Additive Model

\[ g(y) = h_0 + \sum_i h_i(x_i) + \sum_{i \neq j} h_{ij}(x_i, x_j) \]
Chicago Police “Strategic Subject”.

- A risk score for individuals being victims or offenders in a shooting incident
- 16 features
  - 8 reported being used by Chicago Police
Features Reported being Used

green - model being audited
red - mimic model
Features Reported Not Being Used

green - model being audited
red - mimic model
Auditing COMPAS

green - model being audited
red - mimic model
green - model being audited
red - mimic model

\[ g(y) = h_0 + \sum_i h_i(x_i) + \sum_{i \neq j} h_{ij}(x_i, x_j) \]
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
- 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
Reading Assignments (ML Auditing)

- Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, arXiv 2016
- Malgieri, Gianclaudio. The concept of fairness in the GDPR: a linguistic and contextual interpretation, FAccT 2020
- Goodman, Bryce, and Seth Flaxman. European Union regulations on algorithmic decision-making and a “right to explanation”, AI magazine 2017
- Bellamy, Rachel KE, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, arXiv 2018