Fair Visual Representations

May 8, 2020
Dr. Wei Wei, Prof. James Landay

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

- Gender Swapping for Coreference Resolution

Rudinger et al., 2018
Recap

- Debiasing Word Embedding by Gender Attribute Separation

\[ w^{(g)} \text{ of All Occupations} \]
\[ w^{(a)} \text{ of GloVe for Gender Neutral Occupations} \]
\[ w^{(a)} \text{ of Gender-Neutral GloVE for Gender Neutral Occupations} \]

\[ w^{(g)} - \text{Gender-related Components} \]
\[ w^{(a)} - \text{Gender-neutral Components} \]
The Use of Pre-trained NLP Encoders

- Pre-trained Encoders Are Widely Used in NLP
  - Transfer information from a related domain
  - Boost performance on a small data set
  - Trained through a proxy task

- Pre-trained NLP Encoders
  - ELMO (Peters et al, 2018)
  - BERT (Devlin et al., 2018)
  - XLNet (Yang et al., 2019)

- Can Pre-trained Encoders Be Biased?
Adversarial Learning

Elazar et al, 2018
Twitter Prediction Problem

- Twitter Sentiment & Mention Detection
- Protected Attributes
  - Race
  - Gender
  - Age
- Leakage
  - Predict protected attributes

<table>
<thead>
<tr>
<th>Data</th>
<th>Task</th>
<th>Protected Attribute</th>
<th>Balanced Task Acc</th>
<th>Leakage</th>
<th>Unbalanced Task Acc</th>
<th>Leakage</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIAL</td>
<td>Sentiment</td>
<td>Race</td>
<td>67.4</td>
<td>64.5</td>
<td>79.5</td>
<td>73.5</td>
</tr>
<tr>
<td></td>
<td>Mention</td>
<td>Race</td>
<td>81.2</td>
<td>71.5</td>
<td>86.0</td>
<td>73.8</td>
</tr>
<tr>
<td>PAN16</td>
<td>Mention</td>
<td>Gender</td>
<td>77.5</td>
<td>60.1</td>
<td>76.8</td>
<td>64.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age</td>
<td>74.7</td>
<td>59.4</td>
<td>77.5</td>
<td>59.7</td>
</tr>
</tbody>
</table>

Elazar et al, 2018
Main Task and Adversary Accuracies

Elazar et al, 2018
Beefing Up the Adversary

- Increase the Capacity of the Adversary
  - Model Capacity
  - Weight on Loss
  - Ensemble

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>DIAL Sentiment</th>
<th>DIAL Race</th>
<th>Δ</th>
<th>Mention</th>
<th>Gender</th>
<th>PAN16 Mention</th>
<th>PAN16 Age</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Adversary Baseline</td>
<td>-</td>
<td>67.4</td>
<td>14.5</td>
<td>-</td>
<td>77.5</td>
<td>10.1</td>
<td>74.7</td>
<td>9.4</td>
<td>-</td>
</tr>
<tr>
<td>Standard Adversary</td>
<td>(300/1.0/1)</td>
<td>64.7</td>
<td>6.0</td>
<td>5.0</td>
<td>75.6</td>
<td>8.5</td>
<td>72.5</td>
<td>7.3</td>
<td>6.9</td>
</tr>
<tr>
<td>Adv-Capacity</td>
<td>500</td>
<td>64.1</td>
<td>6.7</td>
<td>5.2</td>
<td>73.8</td>
<td>8.1</td>
<td>71.4</td>
<td>4.3</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>63.4</td>
<td>7.1</td>
<td>4.9</td>
<td>75.2</td>
<td>8.9</td>
<td>71.6</td>
<td>6.3</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>65.2</td>
<td>8.1</td>
<td>6.9</td>
<td>76.1</td>
<td>6.7</td>
<td>71.9</td>
<td>6.0</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>63.9</td>
<td>6.2</td>
<td>3.7</td>
<td>74.5</td>
<td>5.6</td>
<td>73.0</td>
<td>10.2</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>8000</td>
<td>65.0</td>
<td>7.1</td>
<td>4.8</td>
<td>75.7</td>
<td>5.4</td>
<td>71.9</td>
<td>9.8</td>
<td>7.3</td>
</tr>
<tr>
<td>λ</td>
<td>0.5</td>
<td>63.9</td>
<td><strong>6.8</strong></td>
<td>6.2</td>
<td>75.6</td>
<td>7.8</td>
<td>73.1</td>
<td>4.8</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>64.9</td>
<td>7.4</td>
<td>5.4</td>
<td>75.6</td>
<td><strong>4.9</strong></td>
<td>72.5</td>
<td>6.8</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>64.2</td>
<td>7.3</td>
<td>5.9</td>
<td>76.0</td>
<td>-7.2</td>
<td>72.1</td>
<td>8.5</td>
<td>7.7</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>65.8</td>
<td>10.2</td>
<td>10.1</td>
<td>73.7</td>
<td>6.4</td>
<td>72.5</td>
<td>-6.3</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>50.0</td>
<td>-</td>
<td>-</td>
<td>73.6</td>
<td>6.5</td>
<td>69.0</td>
<td><strong>3.2</strong></td>
<td>2.9</td>
</tr>
<tr>
<td>Ensemble</td>
<td>2</td>
<td>62.4</td>
<td>7.4</td>
<td>5.4</td>
<td>74.8</td>
<td>6.4</td>
<td>72.8</td>
<td>8.8</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>66.5</td>
<td>6.5</td>
<td>5.0</td>
<td>75.3</td>
<td>4.9</td>
<td>72.1</td>
<td>6.7</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>63.8</td>
<td><strong>4.8</strong></td>
<td>2.6</td>
<td>74.3</td>
<td><strong>4.1</strong></td>
<td>70.1</td>
<td><strong>5.7</strong></td>
<td>5.4</td>
</tr>
</tbody>
</table>

Δ - the difference between the attacker score and the corresponding adversary’s accuracy

Elazar et al, 2018
Outline

- Counterfactual Fairness
- Counterfactual Face Attribution
- Gender Equalized Image Captioning
- Adversarial Removal of Gender Features
Counterfactual Explanation

\[
x' = \arg \min_{x'} \lambda (\hat{f}(x') - y')^2 + d(x, x')
\]

**Counterfactual example**

**Desired outcome**

**Distance function**

Increase \( \lambda \) while \( |\hat{f}(x') - y'| > \varepsilon \)

\[
\begin{align*}
\Delta x &= ? \\
X' &\xrightarrow{f} Y' \\
X &\xrightarrow{f} Y \\
\text{Can remain as black box} &\xrightarrow{Y'} \\
\text{Change to desired outcome}
\end{align*}
\]

Grath et al, 2018
Counterfactual Explanations

Grath et al, 2018
Counterfactual Fairness

\[ P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a) \]

Kusner et al, 2017
Causal View of Counterfactual Fairness

\[ P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) \]

Kusner et al, 2017
Causal View of Counterfactual Fairness

\[ P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a) \]

Kusner et al., 2017
Red Car Problem

- Develop a Fair Algorithm to Determine Insurance Premium

\[
P(y = \text{acc} | A = 1) = 0.2 \quad P(y = \text{acc} | A = 2) = 0.2 \quad P(y = \text{acc} | A = 3) = 0.2
\]

Kusner et al, 2017
Causal Perspective of the Red Car Problem

Observed Variables

Latent (Unobserved) Variables

Kusner et al, 2017

\[ P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a) \]
# Fairness Criteria

<table>
<thead>
<tr>
<th>Individual Treatment</th>
<th>Group Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fairness Through Unawareness</strong></td>
<td><strong>Demographic Parity</strong></td>
</tr>
<tr>
<td>Excludes Sensitive Information A from the predictor</td>
<td>[ P(\hat{Y} = 1</td>
</tr>
<tr>
<td><strong>Individual Fairness</strong></td>
<td><strong>Equal Opportunity/Odds</strong></td>
</tr>
</tbody>
</table>
| \[ M(x_i) \approx M(x_j) | d(x_i, x_j) \approx 0 \] | \[ P(\hat{Y} = 1 | A = 0, Y = 1) = P(\hat{Y} = 1 | A = 1, Y = 1) \]
| | \[ P(\hat{Y} = 1 | A = 0, Y) = P(\hat{Y} = 1 | A = 1, Y) \] |
| **Counterfactual Fairness** | |
| \[ P(\hat{Y}_{A \rightarrow a} (U) = y | X = x, A = a) = P(\hat{Y}_{A \rightarrow a'}(U) = y | X = x, A = a) \] | |

Here, \( M \) represents the function to be satisfied by the predictor, \( x_i \) and \( x_j \) are vectors of features, \( d(x_i, x_j) \) is the distance between \( x_i \) and \( x_j \), and \( A \) is the sensitive information.
Outline

- Counterfactual Fairness
- Counterfactual Face Attribution
- Gender Equalized Image Captioning
- Adversarial Removal of Gender Features
Counterfactual Face Attribution

- Evaluate the Counterfactual Fairness of Face Recognition Systems

\[
P(\hat{Y}_{A \leftarrow a}(U) = y | X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y | X = x, A = a)
\]
CelebA Dataset

Eyeglasses  Wearing Hat

Bangs  Wavy Hair

Pointy Nose  Mustache

Oval Face  Smiling

Liu et al, 2015
Model Architecture

Denton et al, 2019

z ~ p(z)

z = E(x)

z ~ p(z)

x ~ G(z)

y = f(x)

D(z)

Denten et al, 2019

f(x) - smile classifier

G(z) - generator

D(z) - discriminator

E(x) - encoder
Latent Code Attribution

Latent code $z$

Attribute vector

Latent codes corresponding to images with attribute $a$

Latent codes corresponding to images without attribute $a$
Latent Code Manipulation

Denton et al, 2019
Counterfactual Fairness Assessment

\[ P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a) \]

Denton et al, 2019
Sensitivity

\[ S_f(d) = \mathbb{E}_{z \sim p(z)}[f(G(z + d)) - f(G(z))] \]

\[ P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a) \]
Sensitivity Results

\[ S_f(d) = \mathbb{E}_{z \sim p(z)}[f(G(z + d)) - f(G(z))] \]
Heavy Makeup
Young
Directional Sensitivity

\[ S_{y}^{0 \rightarrow 1}(d) = \mathbb{E}_{z \sim p(z) | y(G(z)) = 0} \mathbb{I}[y(G(z + d))! = y(G(z))] \]

from "not smiling" to "smiling"

\[ S_{y}^{1 \rightarrow 0}(d) = \mathbb{E}_{z \sim p(z) | y(G(z)) = 1} \mathbb{I}[y(G(z + d))! = y(G(z))] \]

from "smiling" to "not smiling"

\[ S_f(d) = \mathbb{E}_{z \sim p(z)}[f(G(z + d)) - f(G(z))] \]
### Sensitivity Results

| CelebA attribute defining $d_a$ | $S_{y ightarrow 0}^1(d_a)$ | $S_{y ightarrow 1}^0(d_a)$ |
|---------------------------------|-----------------------------|-----------------------------|
| Young                           | 7.0%                        | 2.6%                        |
| 5_o_Clock_Shadow                | 11.8%                       | 2.2%                        |
| Goatee                          | 12.4%                       | 0.9%                        |
| No_Beard                        | 0.8%                        | 11.8%                       |
| Heavy_Makeup                    | 1.6%                        | 12.4%                       |
| Wearing_Lipstick                | 1.7%                        | 16.3%                       |
Outline

- Counterfactual Fairness
- Counterfactual Face Attribution
- Gender Equalized Image Captioning
- Adversarial Removal of Gender Features
The Equalizer Model

Wrong

Baseline: A man sitting at a desk with a laptop computer.
Our Model: A woman sitting in front of a laptop computer.

Right for the Right Reasons

Baseline: A man holding a tennis racquet on a tennis court.
Our Model: A man holding a tennis racquet on a tennis court.

Right for the Wrong Reasons

Burns et al, 2019
The Basic Idea

Original Image  \( I = \)

Annotated Gender Neutral Image  \( I' = \)

Burns et al, 2019
The Equalizer Model

\[ L = \alpha L^{CE} + \beta L^{AC} + \mu L^{Con} \]

- Cross Entropy Loss
  \[ L^{CE} = -\frac{1}{N} \sum_{n=0}^{N} \sum_{t=0}^{T} \log(p(w_t|w_{0:t-1}, I)) \]
- Appearance Confusing Loss on the gender neutral image
- Confidence Loss on the original image

Burns et al, 2019
Appearance Confusing Objective

\[ \mathcal{L}^{AC} = \frac{1}{N} \sum_{n=0}^{N} \sum_{t=0}^{T} \mathbb{1}(w_t \in G_w \cup G_m) C(\tilde{w}_t, I') \]

\[ C(\tilde{w}_t, I') = \left| \sum_{g_w \in G_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I') - \sum_{g_m \in G_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I') \right| \]

Push Toward Extremes

\[ p(\tilde{w}_t = g_w | w_{0:t-1}, I') \quad \leftrightarrow \quad p(\tilde{w}_t = g_m | w_{0:t-1}, I') \]

\[ G_w \] - set of words for woman

\[ G_m \] - set of words for man

Burns et al, 2019
Confidence Objective

\[ L^{Con} = \frac{1}{N} \sum_{n=0}^{N} \sum_{t=0}^{T} (1(w_t \in G_w) \mathcal{F}^W(\tilde{w}_t, I) + 1(w_t \in G_m) \mathcal{F}^M(\tilde{w}_t, I)) \]

\[ \mathcal{F}^W(\tilde{w}_t, I) = \frac{\sum_{g_m \in G_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I)}{(\sum_{g_w \in G_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I)) + \epsilon} \]

\[ G_w - \text{set of words for woman} \]
\[ G_m - \text{set of words for man} \]

Burns et al, 2019
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MSCOCO-Bias Error</th>
<th>MSCOCO-Bias Ratio Δ</th>
<th>MSCOCO-Balanced Error</th>
<th>MSCOCO-Balanced Ratio Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-FT</td>
<td>12.83</td>
<td>0.15</td>
<td>19.30</td>
<td>0.51</td>
</tr>
<tr>
<td>Balanced</td>
<td>12.85</td>
<td>0.14</td>
<td>18.30</td>
<td>0.47</td>
</tr>
<tr>
<td>UpWeight</td>
<td>13.56</td>
<td>0.08</td>
<td>16.30</td>
<td>0.35</td>
</tr>
<tr>
<td>Equalizer w/o ACL</td>
<td>7.57</td>
<td>0.04</td>
<td>10.10</td>
<td>0.26</td>
</tr>
<tr>
<td>Equalizer w/o Conf</td>
<td>9.62</td>
<td>0.09</td>
<td>13.90</td>
<td>0.40</td>
</tr>
<tr>
<td>Equalizer</td>
<td><strong>7.02</strong></td>
<td><strong>-0.03</strong></td>
<td><strong>8.10</strong></td>
<td><strong>0.13</strong></td>
</tr>
</tbody>
</table>

Baseline-FT - basic LSTM attention model ([Xu et al., 2015](#))

Balanced - resampled dataset to have balanced gender ratio

UpWeight - reweighting

Δ - change to the gender ratio compared to the dataset

[Burns et al., 2019](#)
Results

Baseline-FT: A man walking a dog on a leash.

UpWeight: A man and a dog are in the snow.

Equalizer w/o ACL: A man riding a snowboard down a snow covered slope.

Equalizer: A person walking a dog on a leash.

A woman walking down a street holding an umbrella.

A woman walking down a street holding an umbrella.

A man walking down a street holding an umbrella.

A man walking down a street holding an umbrella.

A man standing in a kitchen preparing food.

A man standing in a kitchen preparing food.

A man standing in a kitchen preparing food.

A man standing in a kitchen preparing food.
Outline

- Counterfactual Fairness
- Counterfactual Face Attribution
- Gender Equalized Image Captioning
- Adversarial Removal of Gender Features
Adversarial Removal of Gender Features

Wang et al, 2019
Model Architecture

\[ \mathcal{L} = \sum_i \beta|x_i - \hat{x}_i| + \mathcal{L}_p(pred(h(\hat{x}_i)), y_i) - \lambda \mathcal{L}_c(c(h(\hat{x}_i)), g_i) \]

Wang et al, 2019
Evaluate Sensitive Information Leakage

- Train an attacker $f(y)$ that reverse engineer the gender information

Data Resampling
\[ \forall y : 1/\alpha < \#(m,y)/\#(w,y) < \alpha \]

Bias Amplification
\[ \Delta = \lambda_M(a) - \lambda_D(a) \]

Wang et al, 2019
Accuracy and Bias Results

Wang et al, 2019

original - no debiasing mask
random - adding random noise
blackout-face - black out using a face detector
blur-sgem - black out using ground truth segmentation
blackout-box - blackout using bounding boxes

Wang et al, 2019
Qualitative Results

COCO Results

imSitu Results

Wang et al, 2019
Summary

● Fairness
  ○ Prevents ML models from biasing toward specific groups when allocating favorable outcomes

● Group Treatments
  ○ Demographic Parity
  ○ Equalized Odds/Opportunity

● Individual Treatments
  ○ Fairness Through Awareness Individual Fairness
  ○ Counterfactual Fairness

● Fairness Techniques
  ○ Pre-processing Methods: Resampling, Reweighting, Optimized-preprocessing
  ○ In-processing Methods: Regularization, Adversarial Learning
  ○ Post-processing Methods: Learning to Defer
Summary

● Fair NLP Methods
  ○ Debiasing Word Embeddings
  ○ Data Augmentation
    ■ Gender Swapping
  ○ Fair Representation for Pre-trained Encoders

● Fair Visual Representations
  ○ Counterfactual Face Attribution
  ○ Gender Equalized Image Captioning
  ○ Adversarial Removal of Gender Features
Reading Assignments

- Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, NeurIPS 2017
- Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints, EMNLP 2017
- Yin, Xi, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data, CVPR 2019
- Singh, Ashudeep, and Thorsten Joachims. Fairness of exposure in rankings, KDD 2017
- Buolamwini, Joy, and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification, FAccT 2018
Next Lecture

Mit-term Project Presentations