Leveraging GANs for fairness evaluations

Emily Denton
Research Scientist, Google Brain
Background

ML Fairness seeks to address *algorithmic unfairness*, with a focus on machine learning systems

Very broad research area!

I will be focusing on one specific component: detecting *undesirable bias in computer vision systems*
Unrepresentative training data can lead to disparities in accuracy for different demographics.
Bias in Computer Vision

Bias in Computer Vision

Social biases embedded in data distribution can be reproduced and/or amplified

[Zhao et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. EMNLP, 2017.]

[Hendricks et al. Women also snowboard: Overcoming bias in captioning models. ECCV, 2018]
Bias in Computer Vision

Human reporting bias can affect annotations

(c) A yellow Vespa parked in a lot with other cars.
(d) A store display that has a lot of bananas on sale.

[Misra et al. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR 2016]
Bias in Computer Vision

Human reporting bias can affect annotations

(c) A yellow Vespa parked in a lot with other cars.

(d) A store display that has a lot of bananas on sale.

“Green bananas”

[Misra et al. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR 2016]
Social biases can affect annotations and propagate through ML systems. 

[Misra et al. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. CVPR 2016]
Bias in Computer Vision

Social biases can affect annotations and propagate through ML system

[Rhue. Racial Influence on Automated Perceptions of Emotions. 2019]
How can GANs help?

High quality photo realistic images

[Karras et al. Progressive growing of gans for improved quality, stability, and variation. ICLR, 2018]
How can GANs help?

High quality photo realistic images

Controllable image synthesis

[Karras et al. Progressive growing of gans for improved quality, stability, and variation. ICLR, 2018]
How can GANs help?

Generative techniques provide tools for testing a classifier’s sensitivity to different image features.

Can answer questions of the form:

*How does the classifier’s output change as some characteristic of the image is systematically varied?*

*Is the classifier sensitive to a characteristic that should be irrelevant for the task?*
GANs can help uncover undesirable bias
GANs can help uncover undesirable bias

\[
P(Smile \mid x)
\]

Manipulate facial hair

\[x \rightarrow \text{CNN} \rightarrow x'
\]
GANs can help uncover undesirable bias

Manipulate facial hair

$x$

$\text{CNN}$

$P(\text{Smile} \mid x)$

$x'$

$\text{CNN}$

$P(\text{Smile} \mid x')$

Did it change?
Can observe the effect on a classifiers of **systematically manipulating factors of variation** in an image.
Can observe the effect on a classifiers of **systematically manipulating factors of variation** in an image.

All else being equal, the presence of facial hair should be irrelevant to the classifier.
Experimental setup

Smiling classifier trained on CelebA (128x128 resolution images)

\[
f(x) = P(\text{Smile} = 1|x) \in (0, 1)
\]

\[
y(x) = \mathbb{I}[P(\text{Smile} = 1|x) \geq c] \in \{0, 1\}
\]
Experimental setup

Smiling classifier trained on CelebA (128x128 resolution images)

\[ f(x) = P(Smile = 1|x) \in (0, 1) \]

\[ y(x) = \mathbb{I}[P(Smile = 1|x) \geq c] \in \{0, 1\} \]

Standard progressive GAN trained to generate 128x128 CelebA images

\[ x = G(z) \quad \text{if} \quad z \sim p(z) \]
Experimental setup

Smiling classifier trained on CelebA (128x128 resolution images)

\[ f(x) = P(Smile = 1|x) \in (0, 1) \]

\[ y(x) = \mathbb{I}[P(Smile = 1|x) \geq c] \in \{0, 1\} \]

Standard progressive GAN trained to generate 128x128 CelebA images

\[ x = G(z) \quad , \quad z \sim p(z) \]

Encoder trained to infer latent codes that generated an images

\[ \tilde{z} = E(x) \]
Attribute vectors

Directions in latent space $\mathcal{Z}$ that manipulate a particular factor of variation in the image.

Latent codes corresponding to images with attribute $a$.

Latent codes corresponding to images without attribute $a$. 
Attribute vectors

We infer attribute vectors using binary CelebA annotations

Eyeglasses = 1

Mustache = 1

Blond_Hair = 1

Eyeglasses = 0

Mustache = 0

Blond_Hair = 0
CelebA attribute vectors

$E(\text{Hat})$ $E(\text{No Hat})$

$E(\text{Mustache})$ $E(\text{No Mustache})$

$\text{Mustache} = 0$ $\text{Mustache} = 1$
A note on CelebA attribute vectors

Many of the attributes are **subjective or ill-defined**

Interpretation of category boundaries is **contingent on the annotators**

The resulting **manipulations reflect how the particular attributes were operationalized and measured** within the CelebA dataset
Manipulating images with CelebA attribute vectors

$G(z)$
Manipulating images with CelebA attribute vectors

\[ d_{\text{young}} \rightarrow G(z) \]
Manipulating images with CelebA attribute vectors
Manipulating images with CelebA attribute vectors

$G(z - d_{\text{Young}})$

$G(z)$

$G(z + d_{\text{Young}})$
Quantifying classifier sensitivity

Model $f$ outputs the probability of a smile being present in the image:

$$f(x) = P(Smile = 1|x) \in (0, 1)$$

Sensitivity of the continuous valued output of $f$ to changes defined by the attribute vector $d$:

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

Difference in classifiers’ output that results from moving in direction $d$ in latent space
Quantifying classifier sensitivity

Given a threshold, $0 \leq c \leq 1$, binary classifications are obtained:

$$y(x) = \mathbb{I}[P(Smile = 1|x) \geq c] \in \{0, 1\}$$

Sensitivity of the discrete classification decision to perturbations along an vector $d$ as:

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

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

Frequency with which classification flips from *smiling* to *not smiling*

Frequency with which classification flips from *not smiling* to *smiling*
Quantifying classifier sensitivity

Given a threshold, $0 \leq c \leq 1$, binary classifications are obtained:

$$y(x) = \mathbb{I}[P(\text{Smile} = 1|x) \geq c] \in \{0, 1\}$$

<table>
<thead>
<tr>
<th>CelebA attribute</th>
<th>$S_{y}^{1\rightarrow0}$</th>
<th>$S_{y}^{0\rightarrow1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>7.0%</td>
<td>2.6%</td>
</tr>
</tbody>
</table>

Sensitivity of the discrete classification decision to perturbations along an vector $d$ as:

$$S_{y}^{1\rightarrow0}(d) = \mathbb{E}_{z \sim p(z)}[y(G(z))=1\mathbb{I}[y(G(z+d))!=y(G(z)))]$$

$$S_{y}^{0\rightarrow1}(d) = \mathbb{E}_{z \sim p(z)}[y(G(z))=0\mathbb{I}[y(G(z+d))!=y(G(z)))]$$
Quantifying classifier sensitivity

Given a threshold, $0 \leq c \leq 1$, binary classifications are obtained:

$$y(x) = \mathbb{I}[P(Smile = 1|x) \geq c] \in \{0, 1\}$$

Sensitivity of the discrete classification decision to perturbations along a vector $d$ as:

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

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

<table>
<thead>
<tr>
<th>CelebA attribute</th>
<th>$S_{y \rightarrow 0}^1$</th>
<th>$S_{y \rightarrow 1}^0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>7.0%</td>
<td>2.6%</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Quantifying classifier sensitivity

Given a threshold, $0 \leq c \leq 1$, binary classifications are obtained:

$$y(x) = \mathbb{I}[P(Smile = 1|x) \geq c] \in \{0, 1\}$$

Sensitivity of the discrete classification decision to perturbations along an vector $d$ as:

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

<table>
<thead>
<tr>
<th>CelebA attribute</th>
<th>$S_{y}^{1 \rightarrow 0}$</th>
<th>$S_{y}^{0 \rightarrow 1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>7.0%</td>
<td>2.6%</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5_o_Clock_Shadow</td>
<td>11.8%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Goatee</td>
<td>12.4%</td>
<td>0.9%</td>
</tr>
<tr>
<td>No_Beard</td>
<td>0.8%</td>
<td>11.8%</td>
</tr>
<tr>
<td>Heavy_Makeup</td>
<td>1.6%</td>
<td>12.4%</td>
</tr>
<tr>
<td>Wearing_Lipstick</td>
<td>1.7%</td>
<td>16.3%</td>
</tr>
</tbody>
</table>
What have the attribute vectors encoded?

~12% of images initially classified as not smiling get classified as smiling after Heavy_Makeup augmentation
What have the attribute vectors encoded?

~12% of images initially classified as not smiling get classified as smiling after Heavy_Makeup augmentation.
What have the attribute vectors encoded?

~7% of images initially classified as smiling get classified as not smiling after Young augmentation
BUT, need to be careful the attribute vector hasn’t actually **encoded** something that should be relevant to smiling classification!

Mouth expression has definitely changed

~40% of images initially classified as not smiling get classified as smiling after High_Cheekbones augmentation
BUT, need to be careful the attribute vector hasn’t actually **encoded** something that should be relevant to smiling classification!

So far we’re verified **makeup, facial hair and age related attribute directions** leave basic mouth shape/smile unchanged

In process of running more of these studies on complete set of attributes
Social context is important

Generative techniques can be used to detect unintended and undesirable bias in facial analysis.

**Equalizing error statistics** across different groups (defined along cultural, demographic, phenotypical lines) is important but not sufficient for building fair, equitable, just or inclusive technology.

This analysis should be part of a larger, socially contextualized, project to critically assess broader ethical concerns relating to facial analysis technology.
Future work

- GAN can be trained on different dataset than classifier
- Increased disentanglement of latent space
- Extend beyond faces
- Other ways of leveraging synthetic data for evaluation (or training?) purposes
  - i.e. mine GANs for data, not people
Related work

**Counterfactual fairness**

**Counterfactual fairness for text**
Garg et al. *Counterfactual Fairness in Text Classification through Robustness*. AIES, 2019

**Individual fairness**

**Model interpretability**
Dabkowski and Gal. *Real time image saliency for black box classifiers*. NIPS, 2017
Thanks!