Google AI

Leveraging GANs for fairness evaluations

Emily Denton
Research Scientist, Google Brain

Emily Denton

Ben Hutchinson

Margaret Mitchell

Timnit Gebru

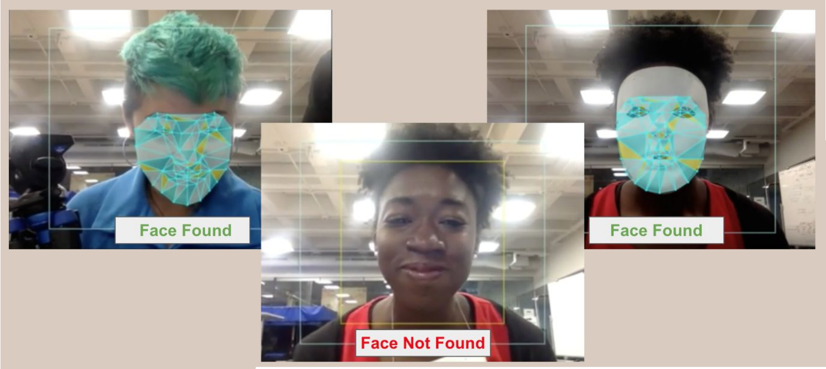
Background

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**

Bias in Computer Vision



The Coded Gaze: Unmasking Algorithmic Bias
Joy Buolamwini

Unrepresentative training data can lead to disparities in accuracy for different demographics

Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products

Inioluwa Deborah Raji
University of Toronto
27 King's College Cir
Toronto, Ontario, Canada, M5S 3H7
deborah.raji@mail.utoronto.com

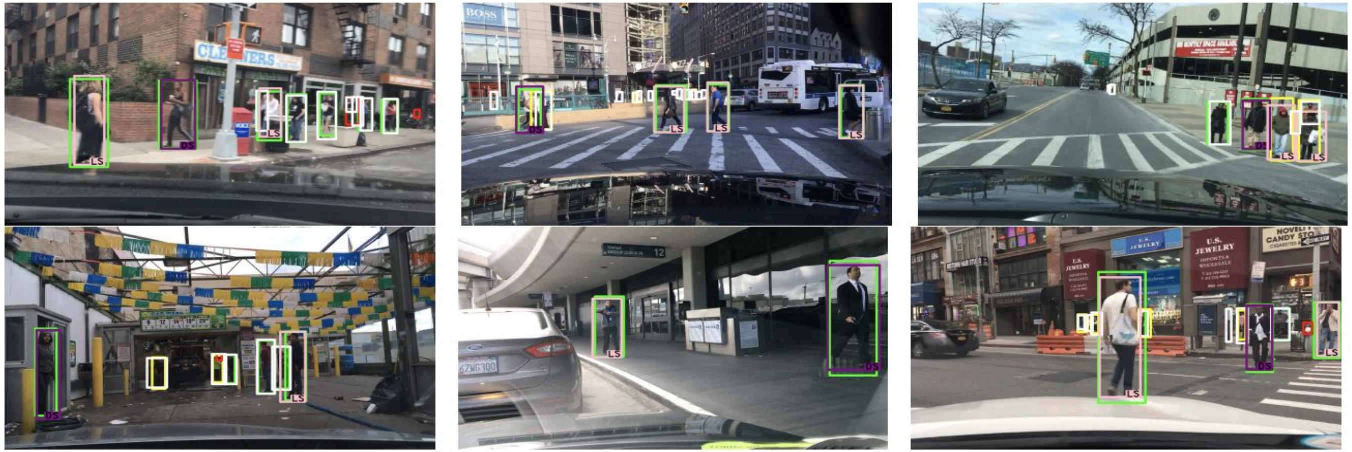
Joy Buolamwini
Massachusetts Institute of Technology
77 Massachusetts Ave
Cambridge, Massachusetts, 02139
joyab@mit.edu

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

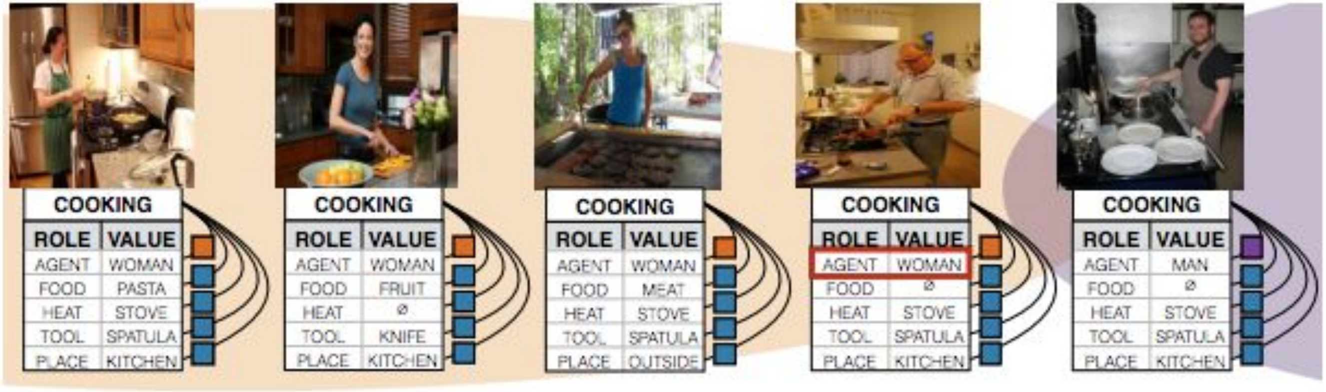
Bias in Computer Vision



[Wilson et al. Predictive inequity in object detection. arXiv:1902.11097, 2019]

Bias in Computer Vision

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

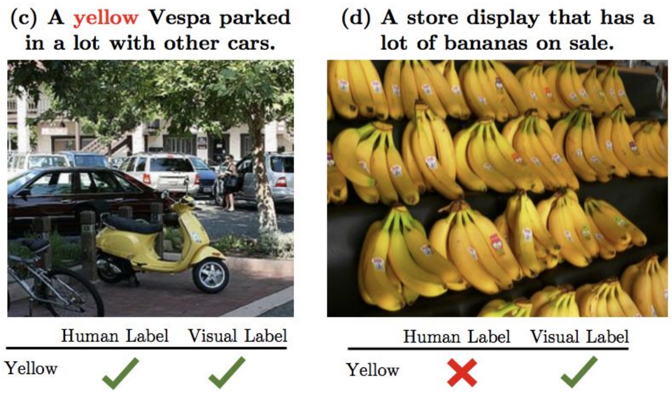


[Zhao et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. EMNLP, 2017.]

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

Bias in Computer Vision

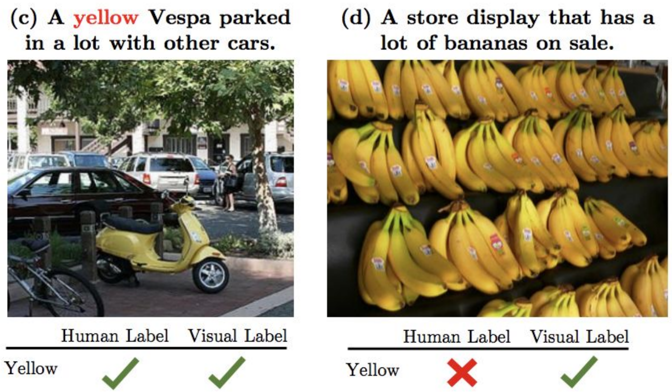
Human reporting bias can affect annotations



[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

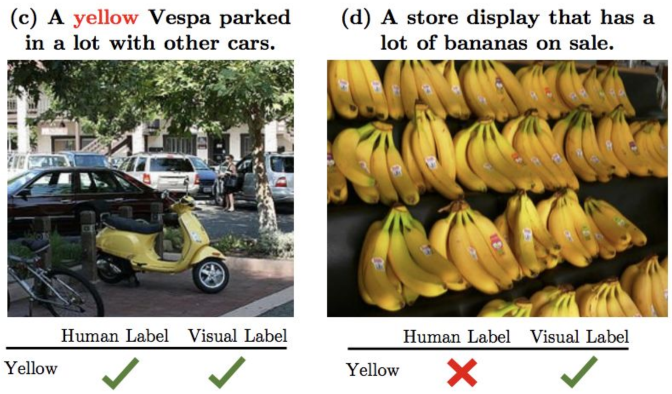


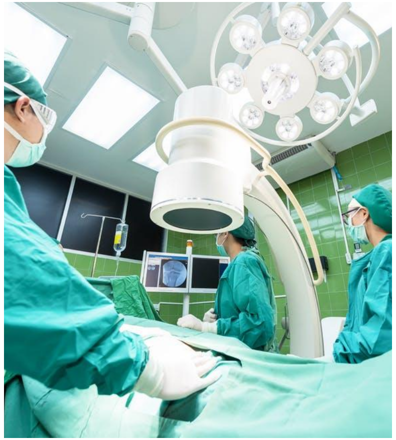
“Green bananas”

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

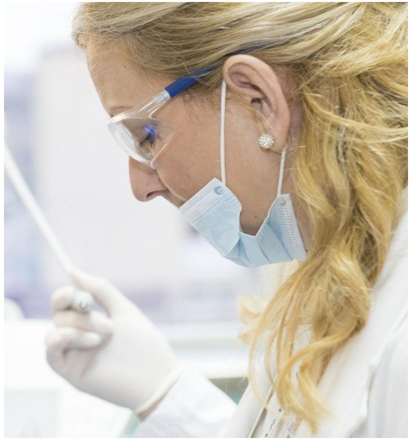
Bias in Computer Vision

Social biases can affect annotations and propagate through ML system





“doctor”



“female doctor”
or
“nurse”

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

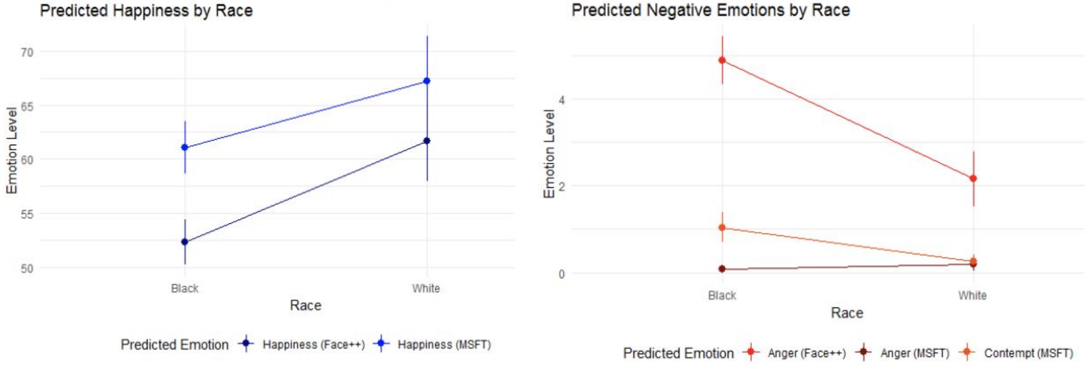
Bias in Computer Vision

Social biases can affect annotations and propagate through ML system

Figure 2. Example Pictures

Figure 2. Darren Collison (L), Gordon Hayward (R) (Source Basketball Reference)

Figure 1. Initial Emotional Comparison



[Rhue. Racial Influence on Automated Perceptions of Emotions. 2019]

How can GANs help?

High quality photo realistic images

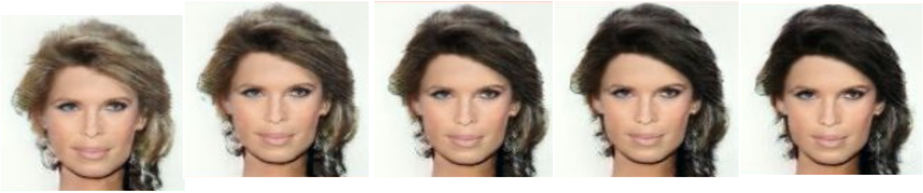
[Karras et al. Progressive growing of gans for improved quality, stability, and variation. ICLR, 2018]

How can GANs help?

High quality photo realistic images

[Karras et al. Progressive growing of gans for improved quality, stability, and variation. ICLR, 2018]

Controllable image synthesis



How can GANs help?

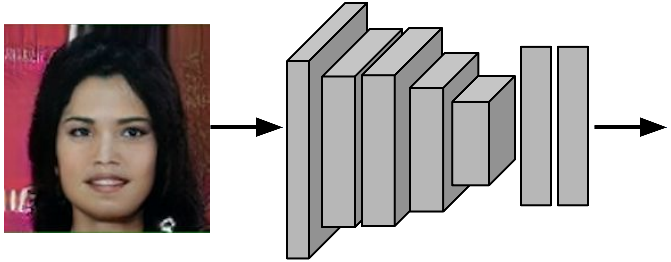
Generative techniques provide tools for testing a classifier's sensitivity to different image features

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

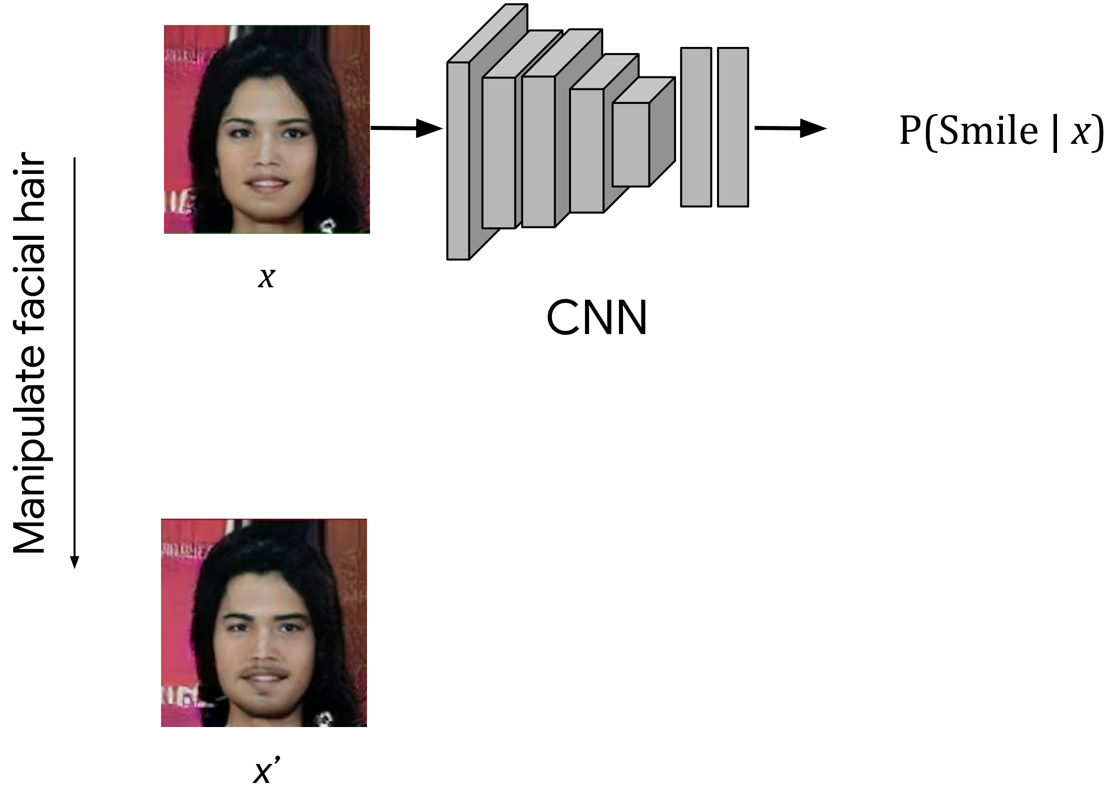


x

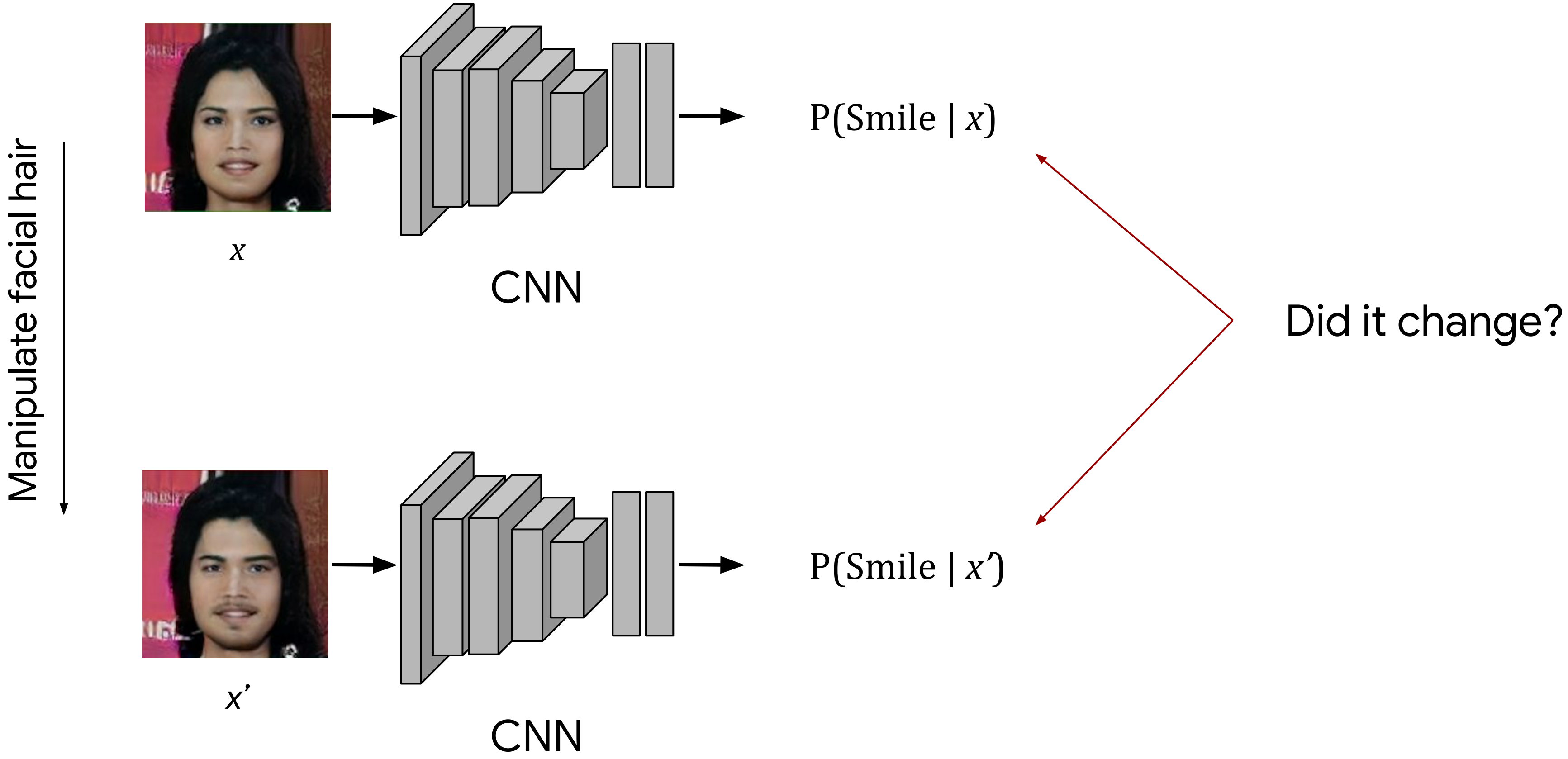
CNN

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

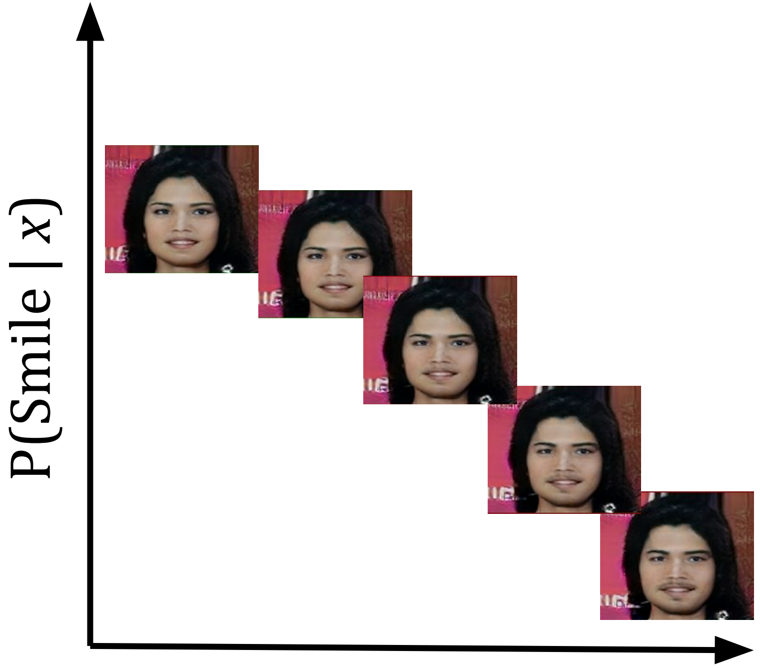
GANs can help uncover undesirable bias

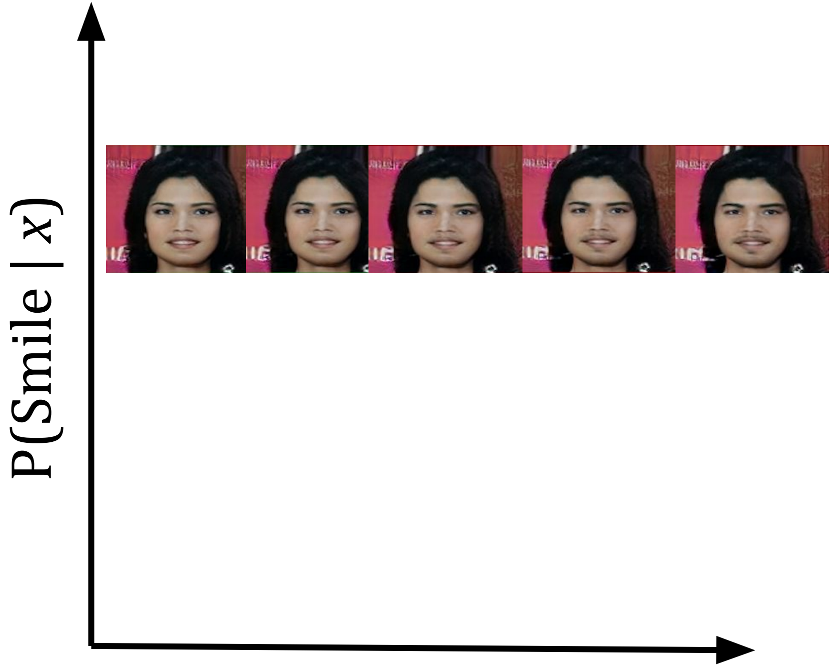


GANs can help uncover undesirable bias

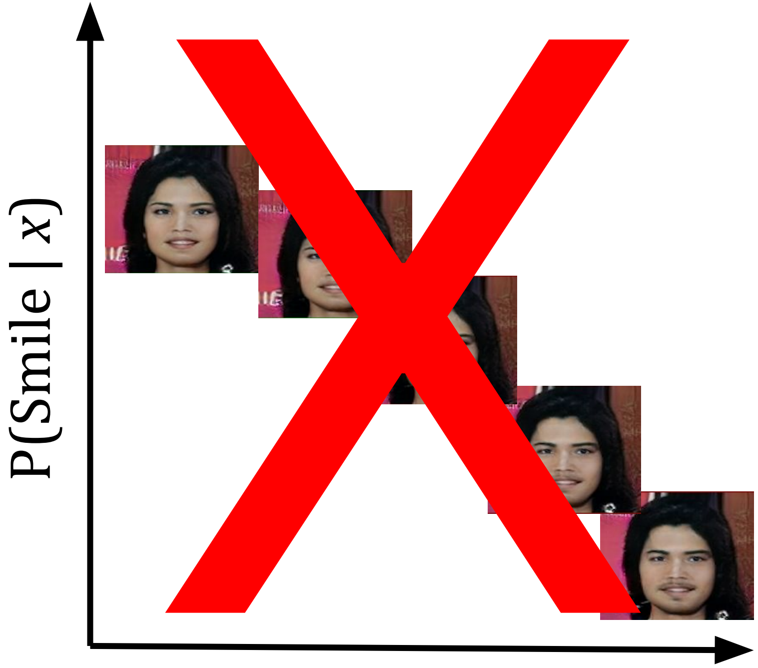


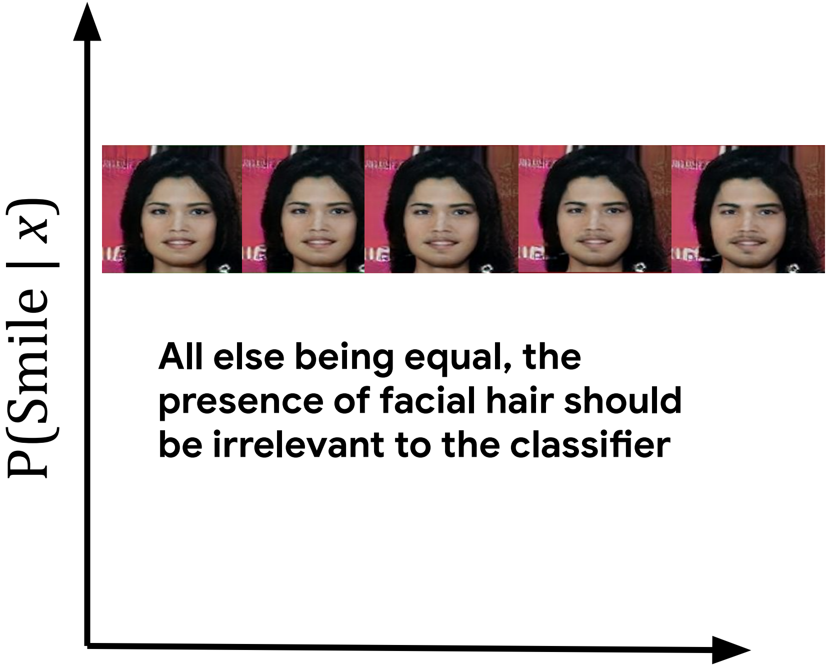
Can observe the effect on a classifiers of **systematically manipulating factors of variation** in an image





Can observe the effect on a classifiers of **systematically manipulating factors of variation** in an image





Experimental setup

Smiling **classifier** trained on CelebA (128x128 resolution images)

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

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

Google

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 $x = G(z) \quad , z \sim p(z)$

Experimental setup

Smiling **classifier** trained on CelebA (128x128 resolution images)

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

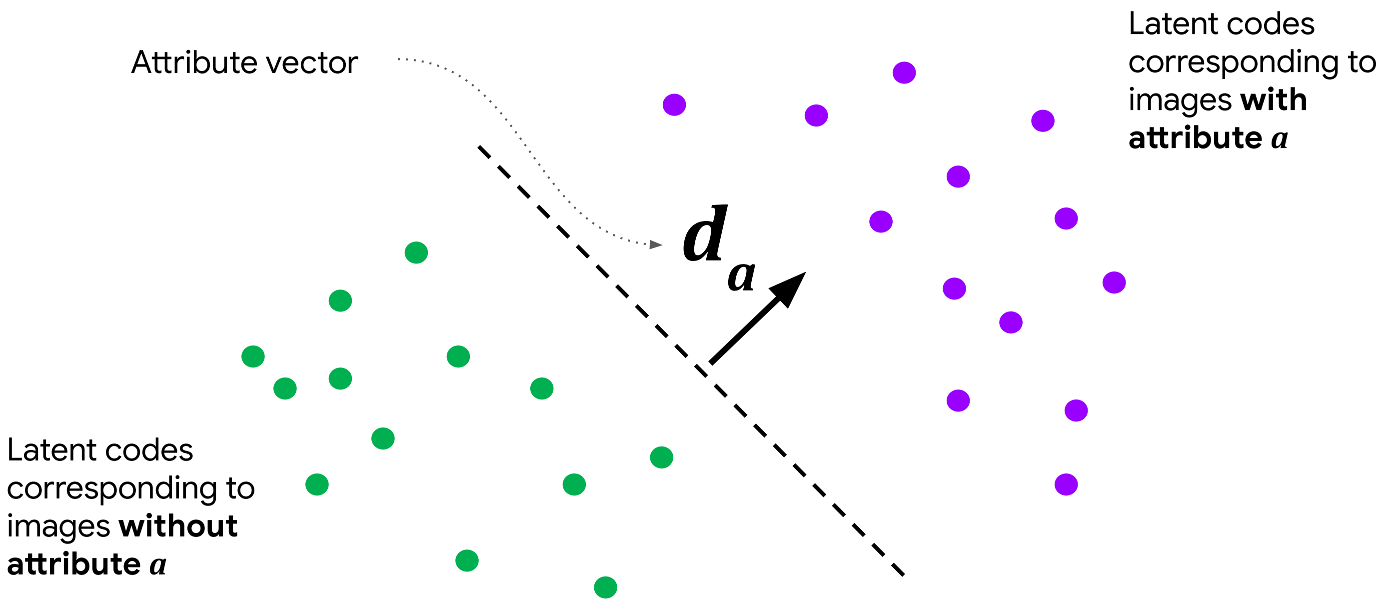
$$y(x) = \mathbb{I}[P(Smile = 1|x) \geq c] \in \{0, 1\}$$

Standard **progressive GAN** trained to generate 128x128 CelebA images $x = G(z) \quad , z \sim p(z)$

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



Attribute vectors

We infer attribute vectors using binary CelebA annotations

Eyeglasses = 1

Eyeglasses = 0

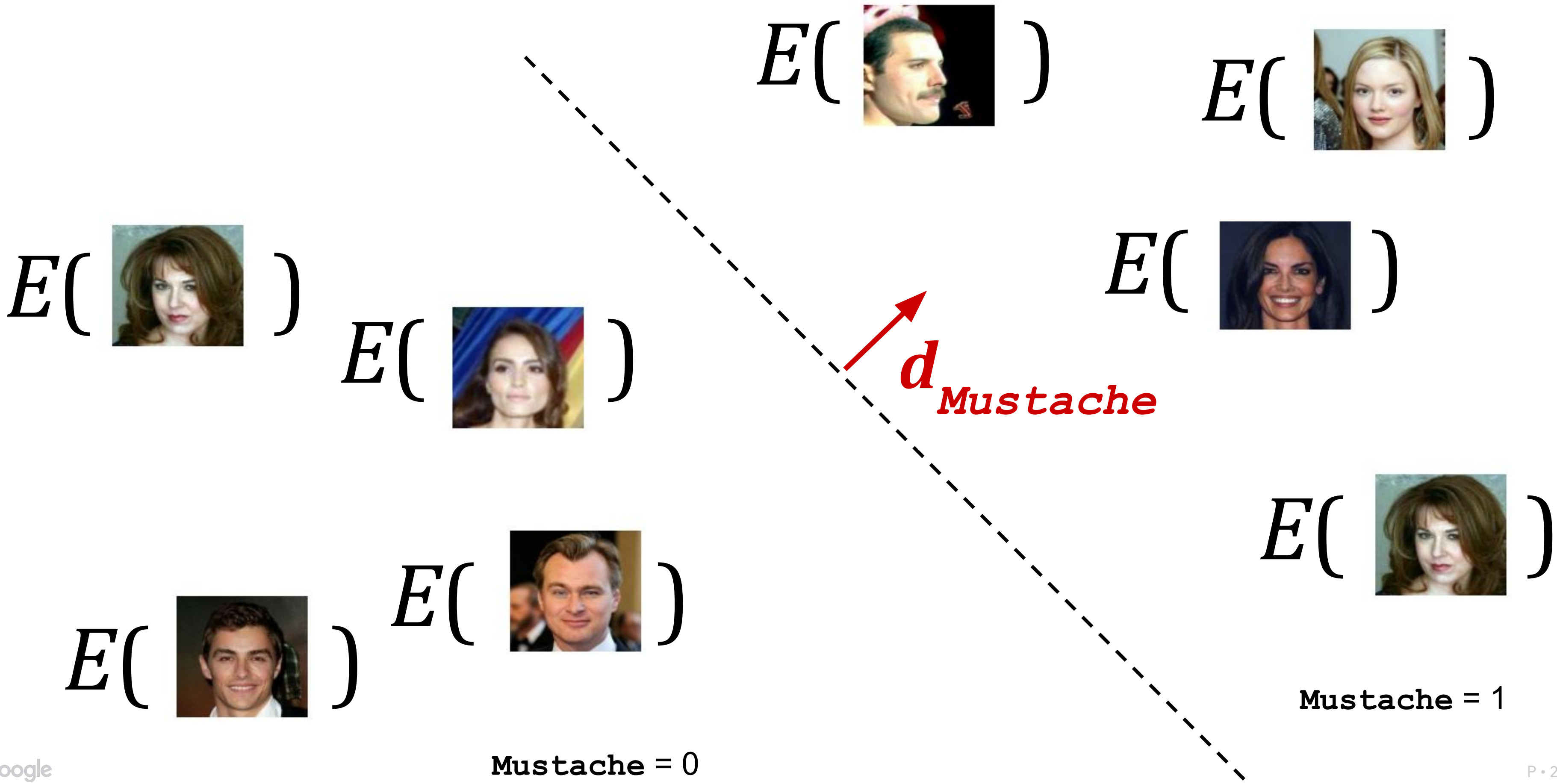
Mustache = 1

Mustache = 0

Blond_Hair = 1

Blond_Hair = 0

CelebA attribute vectors



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

Manipulating images with CelebA attribute vectors

Manipulating images with CelebA attribute vectors

Quantifying classifier sensitivity

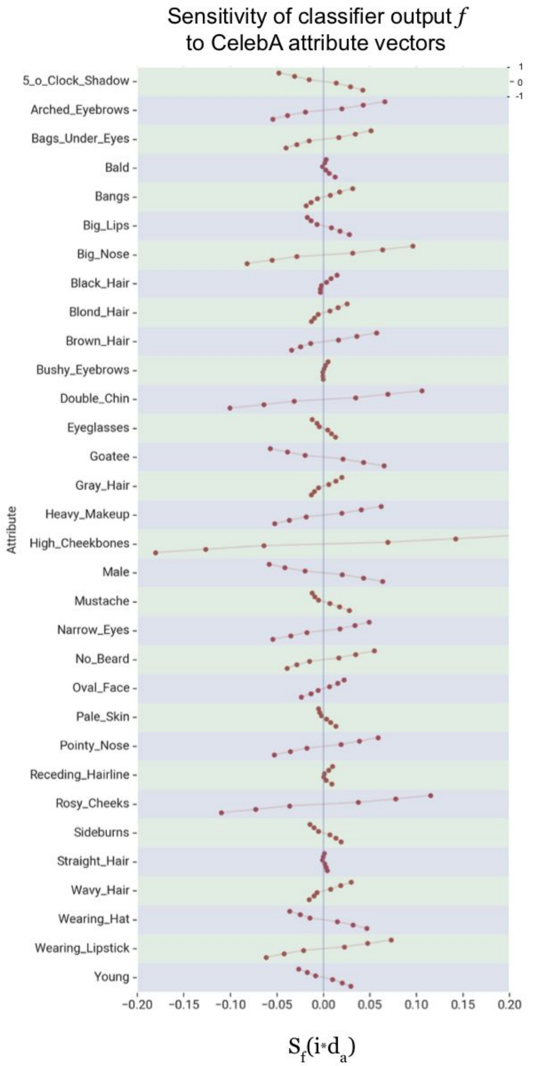
Model f outputs the probability of a smile being present in the image:

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

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

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

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:

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

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

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

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:

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

CelebA attribute	$S_y^{1\to0}$	$S_y^{0\to1}$
Young	**7.0%**	2.6%

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

$$S_y^{1\to0}(d) = \mathbb{E}_{z\sim p(z)|y(G(z))==1}\mathbb{I}[y(G(z+d))! = y(G(z))]$$

$$S_y^{0\to1}(d) = \mathbb{E}_{z\sim p(z)|y(G(z))==0}\mathbb{I}[y(G(z+d))! = y(G(z))]$$

Quantifying classifier sensitivity

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

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

CelebA attribute	$S_y^{1\to 0}$	$S_y^{0\to 1}$
Young	**7.0%**	2.6%
Male		

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

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

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

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 an vector d as:

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

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

CelebA attribute	$S_y^{1\to 0}$	$S_y^{0\to 1}$
Young	**7.0%**	2.6%
Male		
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%**

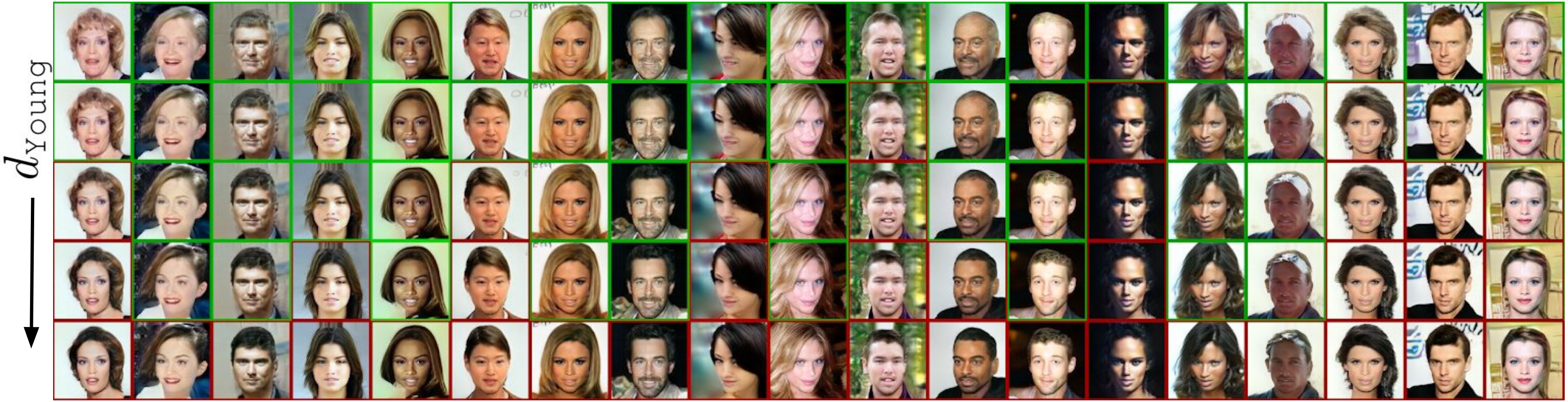
What have the attribute vectors encoded?

~12% of images initially classified as not smiling get classified as smiling after `Heavy_Makeup` augmentation

What have the attribute vectors encoded?

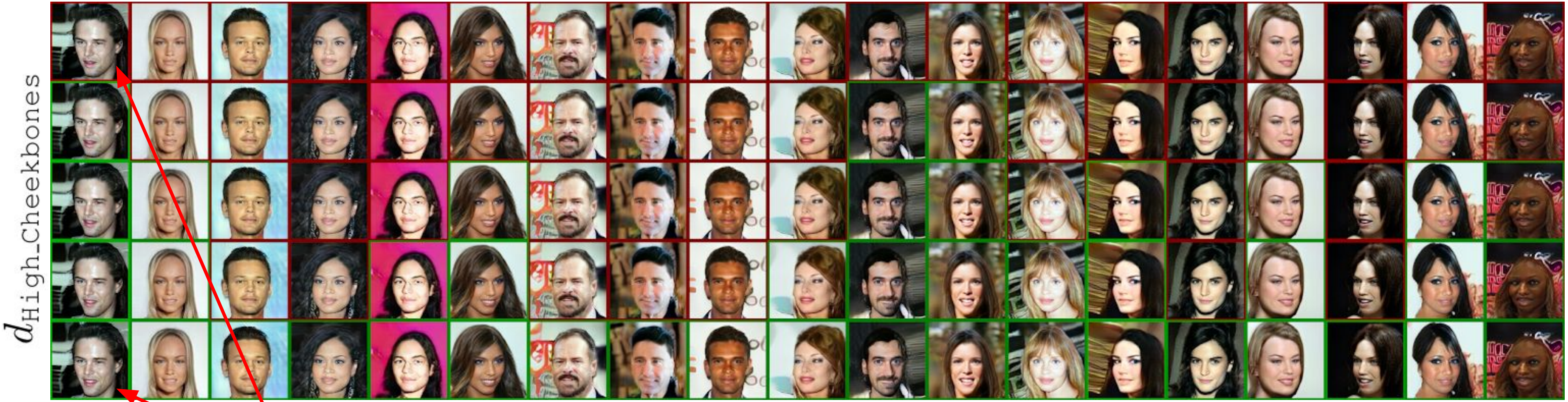
~12% of images initially classified as not smiling get classified as smiling after `Heavy_Makeup` augmentation

What have the attribute vectors encoded?



~7% of images initially classified as smiling get classified as not smiling after `Young` augmentation

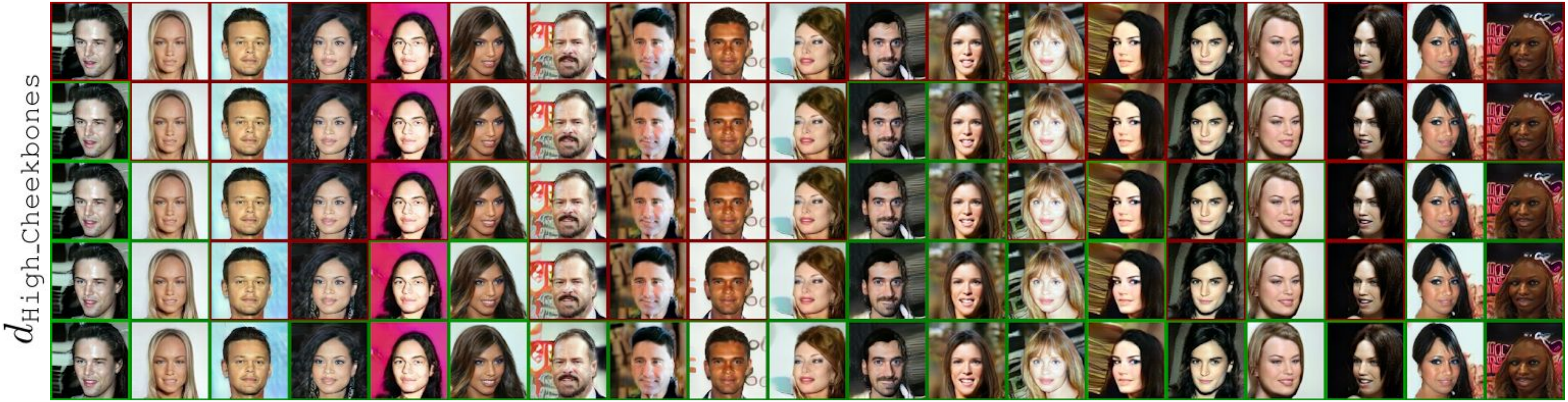
BUT, need to be careful the attribute vector hasn't actually **encoded something that should be relevant to smiling classification**!



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
Kilbertus et al. *Avoiding discrimination through causal reasoning*. NIPS, 2017.
Kusner et al. *Counterfactual fairness*. NIPS, 2017.

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

Individual fairness
Dwork et al. *Fairness Through Awareness*. ITCS, 2012.

Model interpretability
Kim et al. *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)*. ICML, 2018.
Chang et al. *Explaining image classifiers by counterfactual generation*. ICLR, 2019.
Fong and Vedaldi. *Interpretable explanations of black boxes by meaningful perturbation. ICCV*, 2017.
Dabkowski and Gal. *Real time image saliency for black box classifiers*. NIPS, 2017
Simonyan et al. *Deep inside convolutional networks: Visualising image classification models and saliency maps*. 2013

Thanks!

Denton et al. Detecting Bias with Generative Counterfactual Face Attribute Augmentation. *CVPR Workshop on Fairness, Accountability, Transparency and Ethics in Computer Vision*, 2019.