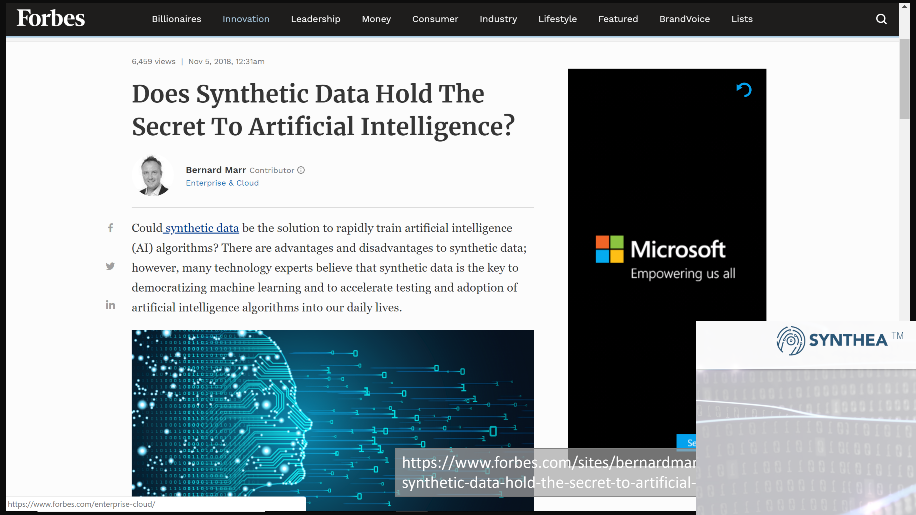
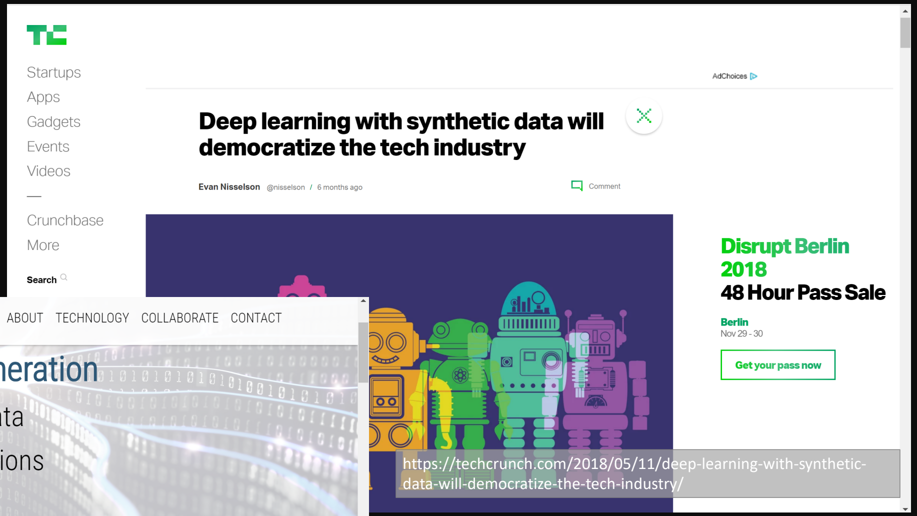
A Study on Generative Adversarial Networks Exacerbating Social Data Bias

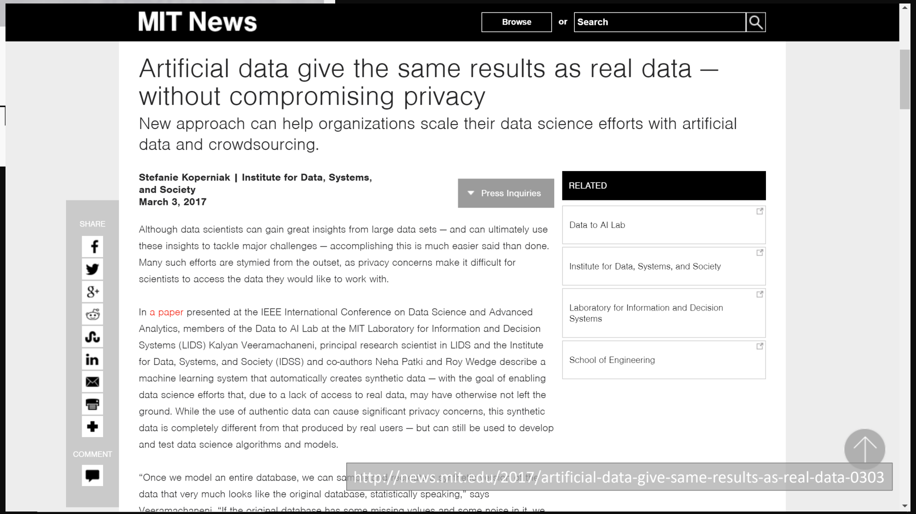
Thesis by Niharika Jain

Chair: Dr. Subbarao Kambhampati

Committee Members: Dr. Huan Liu and Dr. Lydia Manikonda







Machine learning practitioners have celebrated Generative Adversarial Networkss as an economical technique to augment their training sets for data-hungry models when acquiring real data is expensive or infeasible.

It’s not clear that they realize the dangers of this approach!

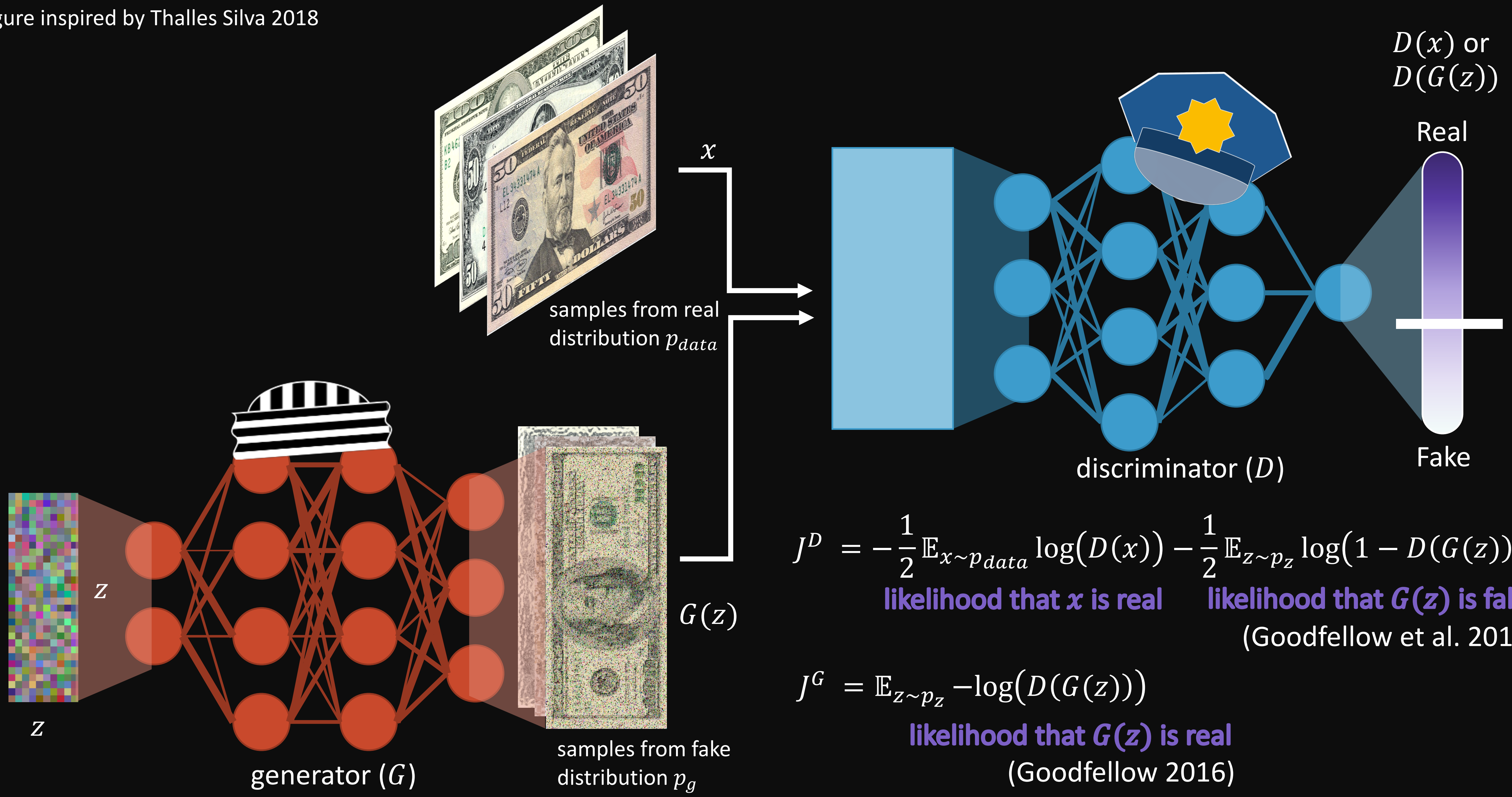
data augmentation

If GANs worked perfectly, they would capture the distribution of the data, and thus capture any biases within it.

GANs have a failure mode which causes them to *exacerbate* bias.

Generative Adversarial Networks: counterfeiter and cop

Figure inspired by Thalles Silva 2018



$$J^D = -\frac{1}{2}\mathbb{E}_{x\sim p_{data}}\log\big(D(x)\big) - \frac{1}{2}\mathbb{E}_{z\sim p_z}\log\big(1 - D(G(z))\big)$$

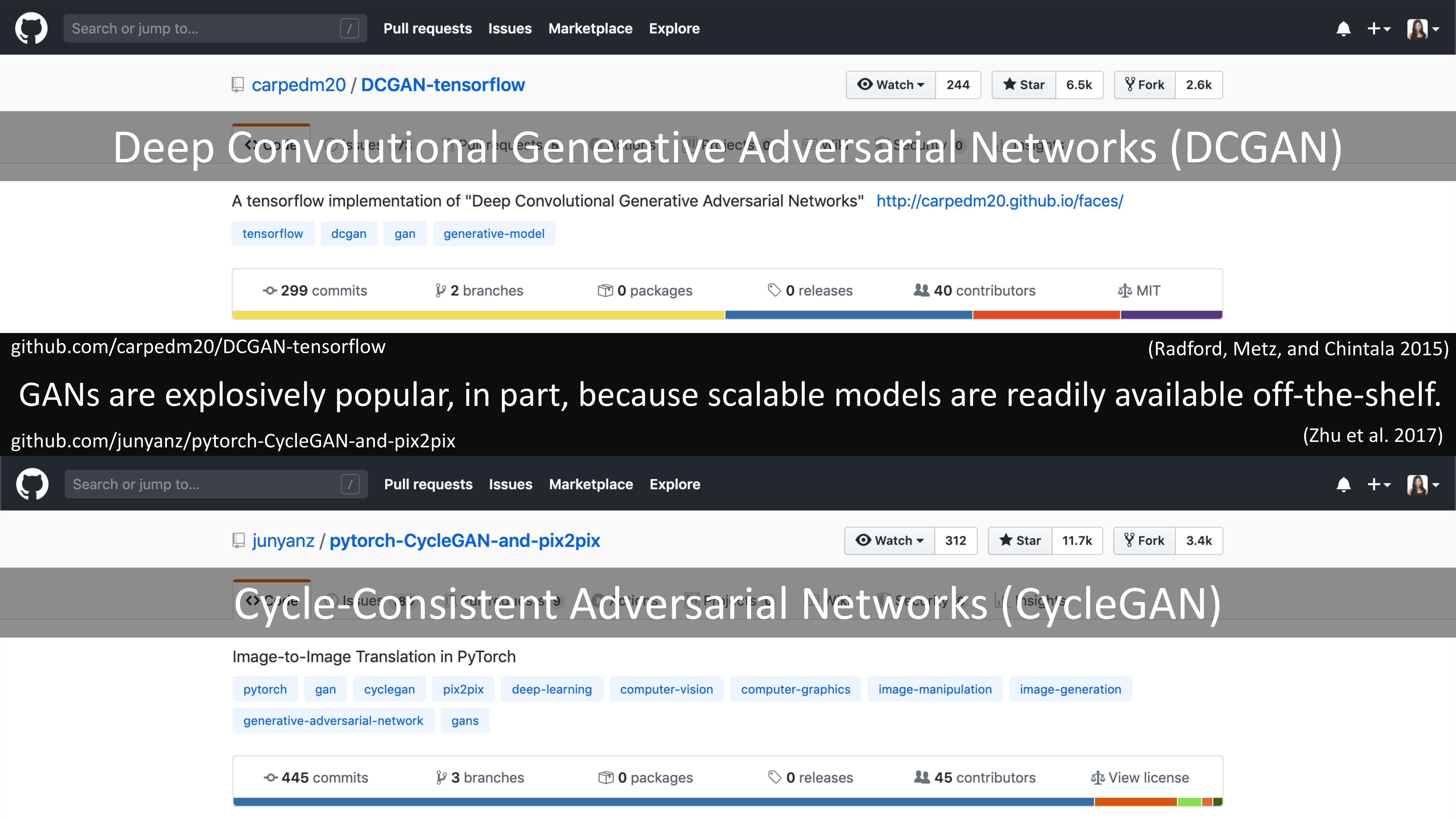
likelihood that x is real — likelihood that $G(z)$ is fake

(Goodfellow et al. 2014)

$$J^G = \mathbb{E}_{z\sim p_z} -\log\big(D(G(z))\big)$$

likelihood that $G(z)$ is real

(Goodfellow 2016)



Search or jump to...
Pull requests
Issues
Marketplace
Explore
carpedm20 / DCGAN-tensorflow
Watch 244
Star 6.5k
Fork 2.6k
Deep Convolutional Generative Adversarial Networks (DCGAN)
A tensorflow implementation of "Deep Convolutional Generative Adversarial Networks" http://carpedm20.github.io/faces/
tensorflow
dcgan
gan
generative-model
299 commits
2 branches
0 packages
0 releases
40 contributors
MIT
github.com/carpedm20/DCGAN-tensorflow
(Radford, Metz, and Chintala 2015)
GANs are explosively popular, in part, because scalable models are readily available off-the-shelf.
github.com/junyanz/pytorch-CycleGAN-and-pix2pix
(Zhu et al. 2017)
Search or jump to...
Pull requests
Issues
Marketplace
Explore
junyanz / pytorch-CycleGAN-and-pix2pix
Watch 312
Star 11.7k
Fork 3.4k
Cycle-Consistent Adversarial Networks (CycleGAN)
Image-to-Image Translation in PyTorch
pytorch
gan
cyclegan
pix2pix
deep-learning
computer-vision
computer-graphics
image-manipulation
image-generation
generative-adversarial-network
gans
445 commits
3 branches
0 packages
0 releases
45 contributors
View license


What do these images have in common?
These are GAN-generated faces, trained on a dataset of engineering professors.

hypothesis:

when a feature is biased in the training set, a GAN amplifies the biases along that dimension in its generated distribution

all biases are equal,
but some are more equal than others.

This hypothesis makes a blanket claim about GANs indiscriminately picking up all types of biases that can exist in the data. For facial images, these biased features could be lighting, facial expression, accessories, or hairstyle.

We only aim to bring attention to exacerbation of *sensitive* features: social characteristics that have been historically discriminated against. This work investigates bias over race and gender.

hypothesis:

when a feature is biased in the training set, a GAN amplifies the biases along that dimension in its generated distribution

for facial datasets, these datasets are often skewed along race and gender, so GANs exacerbate sensitive social biases

don’t try this at home!

Using photos to measure human characteristics has a complicated and dangerous history: in the 19th century, “photography helped to animate—and lend a ‘scientific’ veneer to—various forms of phrenology, physiognomy, and eugenics.” (Crawford and Paglen 2019)

Neither gender nor race can be ascertained from appearance. We use human annotators to classify masculinity of features and lightness of skin color as a crude metric of gender and race to illustrate our argument.

This work is not advocating for the use of facial data in machine learning applications. We create a hypothetical experiment using data with easily-detectable biases to tell a cautionary tale about the shortcomings of this approach.

imagining an engineer

if we train a GAN to imagine faces of US university engineering professors, will it skew the new data toward white males?

We scrape from engineering faculty directories from 47 universities on the U.S. News “Best Engineering Schools” list, remove all noisy images, and crop to the face.

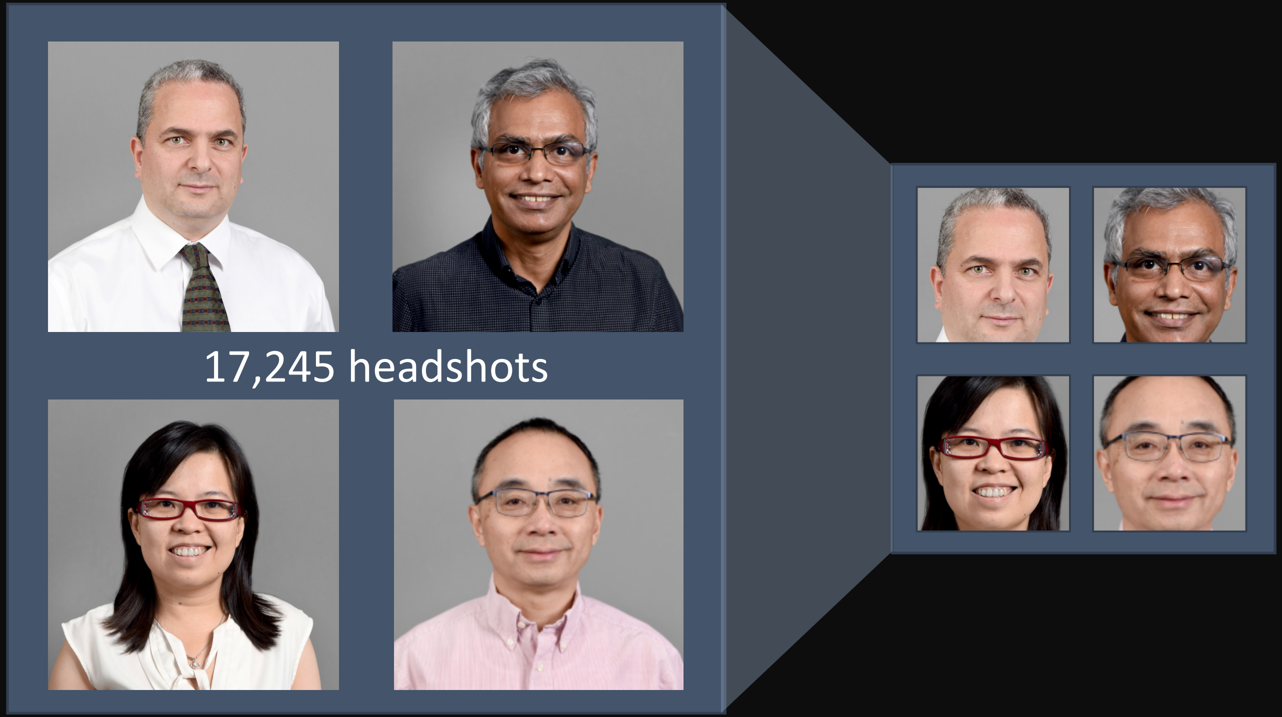
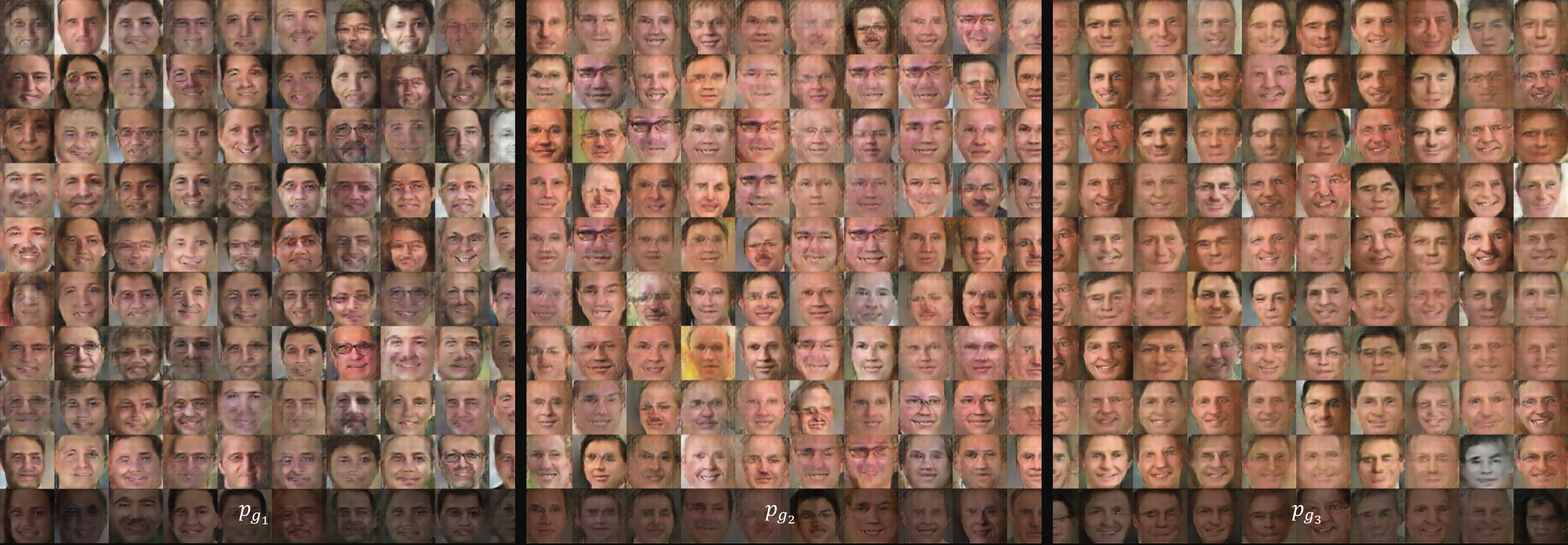


image pre-processing contribution: Alberto Olmo

Images from cidse.engineering.asu.edu/faculty/



DCGAN trained on three random initializations

GAN training contribution: Alberto Olmo

To measure the distributions in their diversity along gender and race, we ask humans on Amazon Mechanical Turk to annotate the images.

For each task, we ask master Turkers to annotate 50 images:

T1a gender on professor images randomly sampled from p_{data}

T1b gender on DCGAN-generated images randomly sampled from p_g

T2a race on professor images randomly sampled from p_{data}

T2b race on DCGAN-generated images randomly sampled from p_g

evaluation

human annotation contribution: Sailik Sengupta

For each image, select the most appropriate description:

- ○ face has mostly masculine features
- ○ face has mostly feminine features
- ✓ neither of the above is true

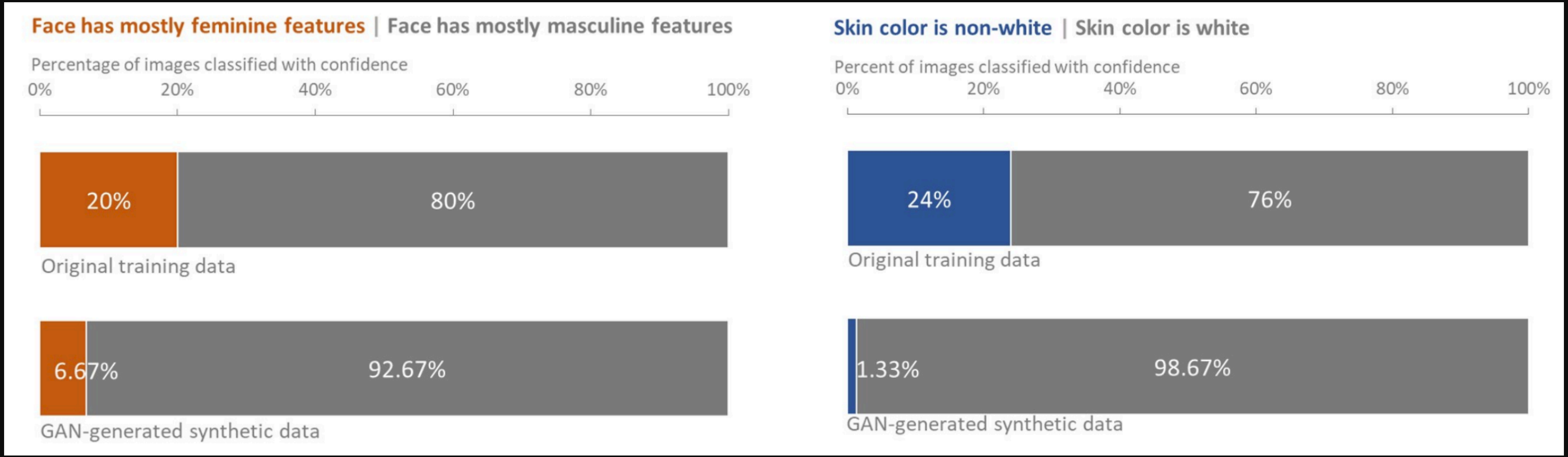
- ✓ skin color is white
- ○ skin color is non-white
- ○ can't tell

Between-subject design: for each distribution ($p_{data}, p_{g_1}, p_{g_2}$, or p_{g_3}), we ask a Turker to annotate 50 images for race and gender.

human annotation contribution: Sailik Sengupta

One-tailed two-proportion z-test

$$H_0\colon \hat{p} = p_0$$
$$H_a\colon \hat{p} < p_0$$

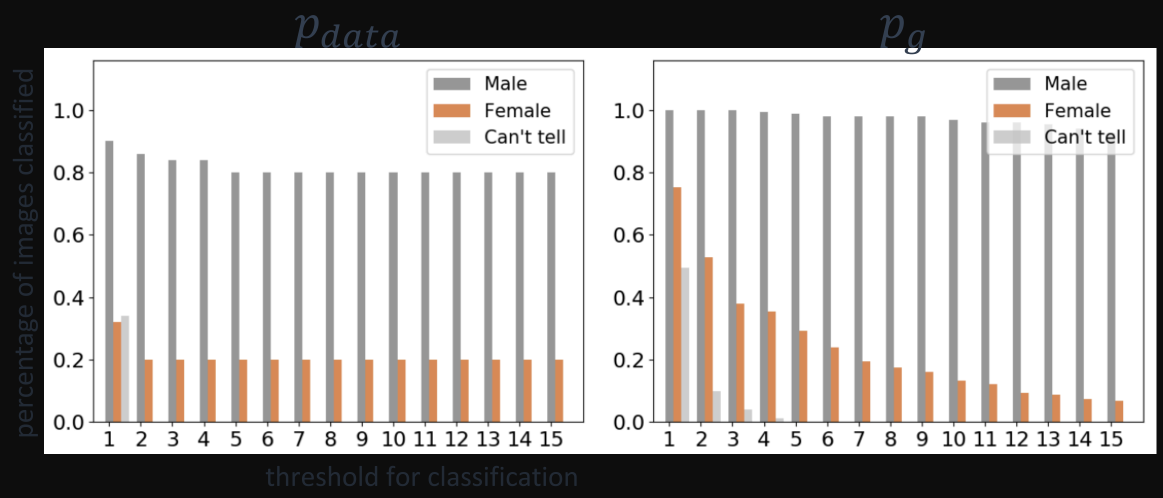


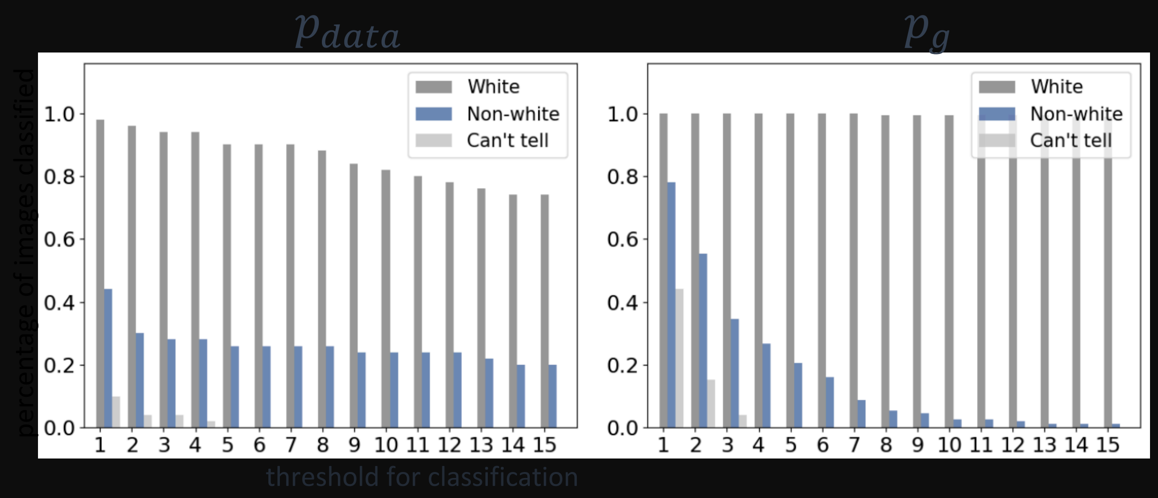
p = 0.0094

p = 0.000087

Using majority thresholding to label images, we find that the representation of minorities is further decreased in the synthetic data.

confidence metrics





gender race

Turkers are not as confident when generated images belong to minority classes as they are when the images belong to the majority. Is human or machine bias to blame?

confidence metrics contribution: Alberto Olmo, Lydia Manikonda