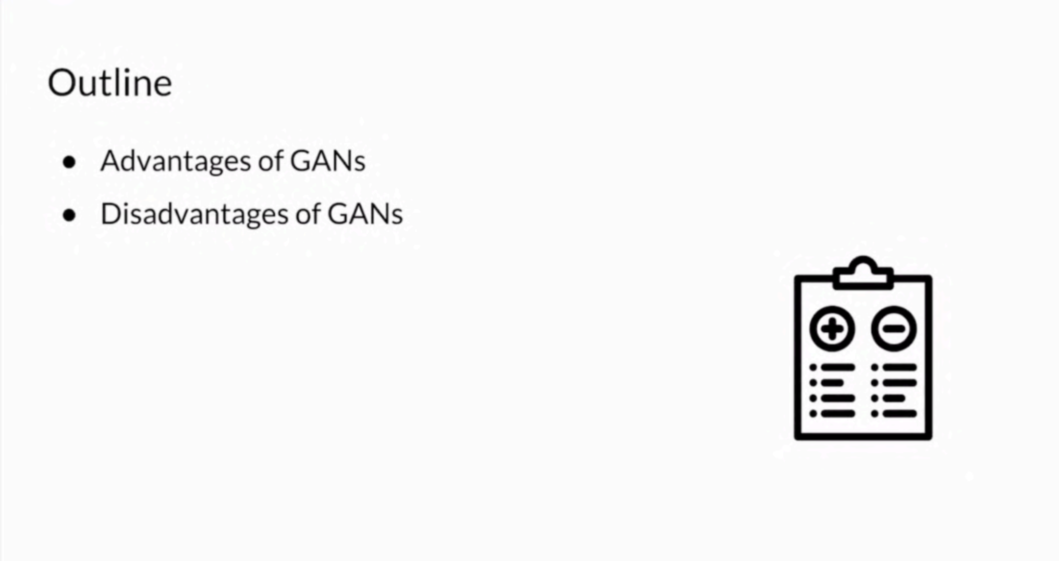
deeplearning.ai

Disadvantages of GANs



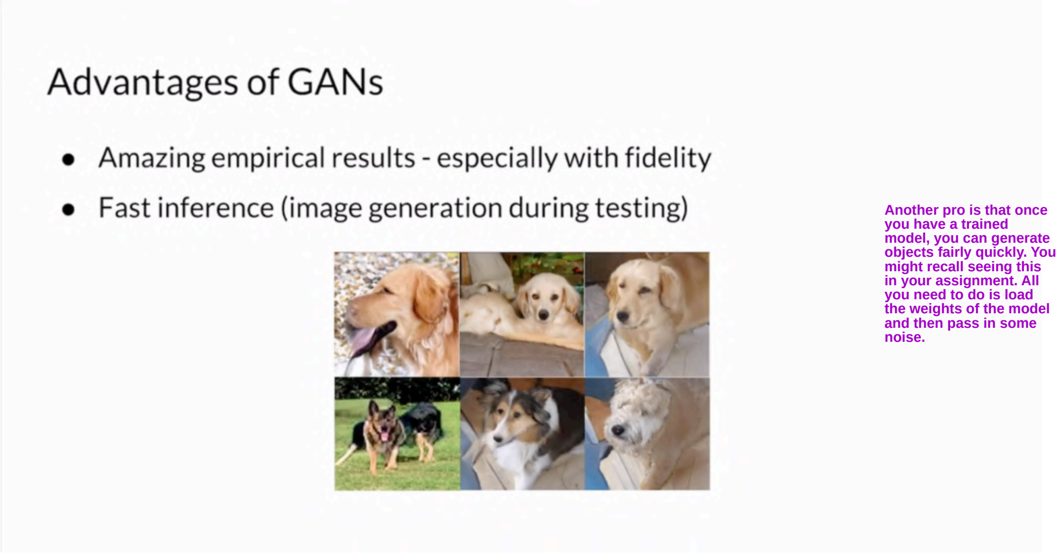
Outline

- Advantages of GANs
- Disadvantages of GANs

Advantages of GANs

- Amazing empirical results - especially with fidelity

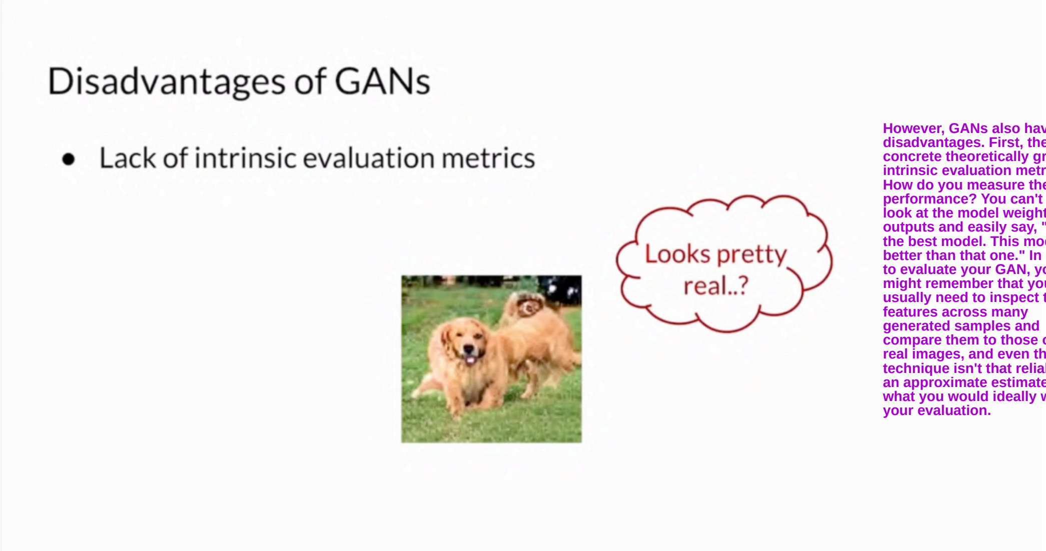
Remember that a significant advantage with GANs is that they can produce high-quality, realistic results. To the human eye like yours and mine, you could be fooled into believing these people actually exist, but these are all generated.



Advantages of GANs

- Amazing empirical results - especially with fidelity
- Fast inference (image generation during testing)

Another pro is that once you have a trained model, you can generate objects fairly quickly. You might recall seeing this in your assignment. All you need to do is load the weights of the model and then pass in some noise.

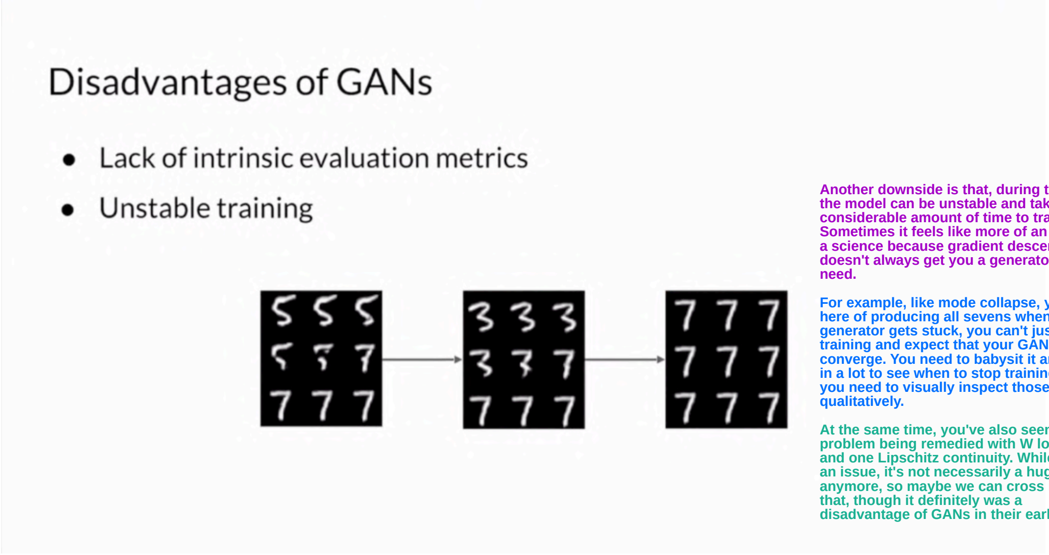


Disadvantages of GANs

- Lack of intrinsic evaluation metrics

Looks pretty real..?

However, GANs also have their disadvantages. First, they lack concrete theoretically grounded intrinsic evaluation metrics. How do you measure their performance? You can't just look at the model weights or outputs and easily say, "This is the best model. This model is better than that one." In order to evaluate your GAN, you might remember that you usually need to inspect the features across many generated samples and compare them to those of the real images, and even that technique isn't that reliable. It's an approximate estimate of what you would ideally want for your evaluation.



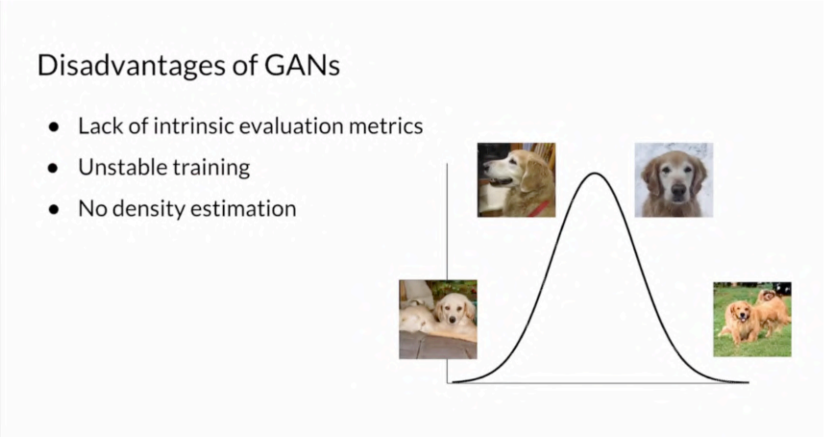
Disadvantages of GANs

- Lack of intrinsic evaluation metrics
- Unstable training

Another downside is that, during training, the model can be unstable and take considerable amount of time to train. Sometimes it feels like more of an art than a science because gradient descent doesn't always get you a generator you need.

For example, like mode collapse, you see here of producing all sevens when the generator gets stuck, you can't just keep training and expect that your GAN will converge. You need to babysit it and check in a lot to see when to stop training, and you need to visually inspect those samples qualitatively.

At the same time, you've also seen this problem being remedied with W loss a bit and one Lipschitz continuity. While this is an issue, it's not necessarily a huge one anymore, so maybe we can cross it out like that, though it definitely was a disadvantage of GANs in their early days.



Disadvantages of GANs

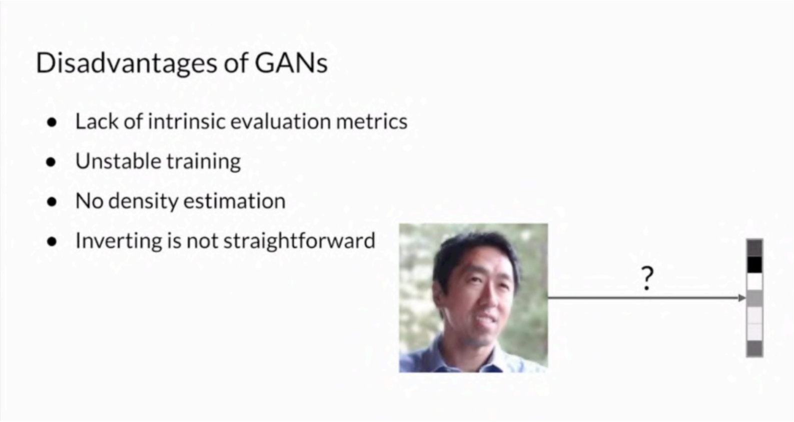
- Lack of intrinsic evaluation metrics
- Unstable training
- No density estimation

Depending on what you want to use your generative model for, GANs might not be the right type of model if you want to explicitly get the probability density over your modeled features, and what that means is, the likelihood of say, a particular image here. How likely are these features to present themselves?

This might be useful if you want to say do anomaly detection, by seeing what an unlikely dog would look like versus a likely dog. Perhaps this is not likely, or it could detect cat dogs that are very unlikely out there.

This is known as density estimation because it's estimating this probability density of all these features.

Density estimation is useful to know how often this golden fur or floppy ears, for example, typically make up a dog, and that can then feed into downstream tasks like finding anomalies out where there's low probability for certain features. Over lots of samples, you could of course get some approximation for your GAN.



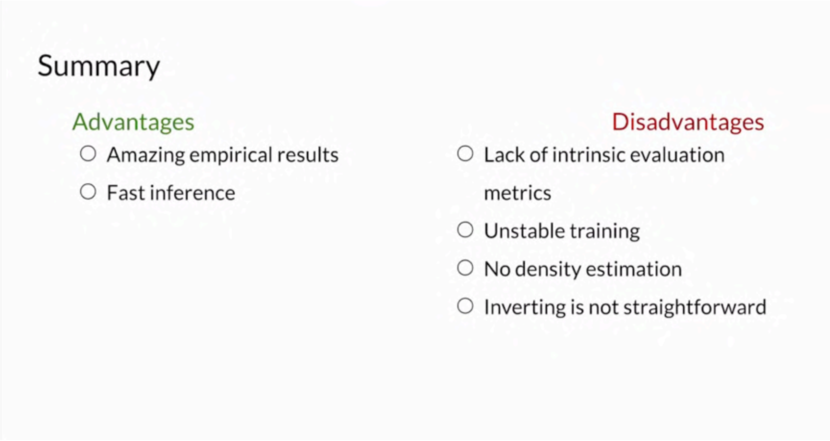
Disadvantages of GANs

- Lack of intrinsic evaluation metrics
- Unstable training
- No density estimation
- Inverting is not straightforward

Lastly, the generator is not trained to be invertible. What that means is that you can take an image such as this one, and be able to figure out what noise vector it maps onto, so the opposite task. Instead of the usual task of inputting a noise factor and then outputting an image. Now you want to feed in an image to figure out what its associated noise vector is.

There have been new methods that have emerged to remedy this problem of invertibility, typically with another model that does the opposite of the GAN, and there are also GANs that are designed to learn both directions at once. One GAN going in one direction and the other one going in the other.

You might be wondering why inversion can be useful, and inversion can be particularly convenient for image editing because that means you can apply your controllable generation skills to that noise vector that you find for any image, and this could be a real image. It doesn't have to be generated already to find that noise vector. Then you tweak that noise vector using those controllable generation skills that you have now, so that this image could be, for example, younger, older, or have blue hair.



Summary

Advantages

○ Amazing empirical results

○ Fast inference

Disadvantages

○ Lack of intrinsic evaluation metrics

○ Unstable training

○ No density estimation

○ Inverting is not straightforward

In summary, GANs have incredibly high quality results and relatively fast generation from a trained model.

However, they lack intrinsic evaluation metrics, have unstable training, though that's been fairly remedied. No formal density estimation that's inherent to the model, and it can be challenging to invert an image to its latent space representation, especially if the model is very large and it's hard to find where that latent might be.

It is important, I think, to emphasize the significance of having high-fidelity results. GANs are arguably the best and arguably the first AI model to achieve such realistic outputs, and very consistently too. GANs are often used elsewhere just to enhance the output's realism. That's really critical to know, and that's where GANs can be applied in so many different areas.

deeplearning.ai

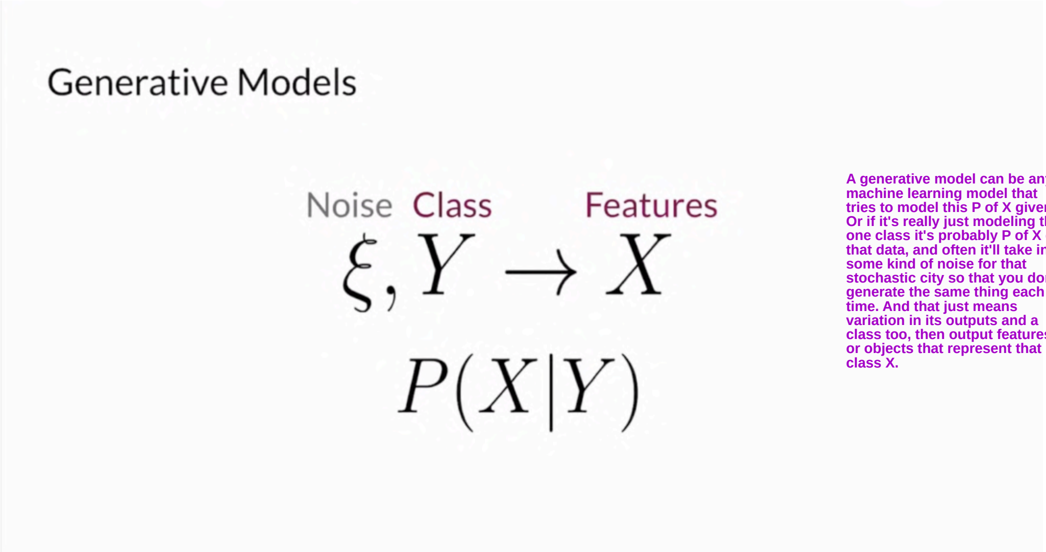
Alternatives to GANs

In the previous section you saw some of the disadvantages to using GANs, in this one you'll see how other generative models address these downsides, but have different trade-offs of their own.

Outline

- Overview of generative models
- VAEs and other alternatives

Specifically, I'll discuss another popular model called VAE that you might already be a little bit familiar with from previous weeks, and then other less popular but still very cool alternatives



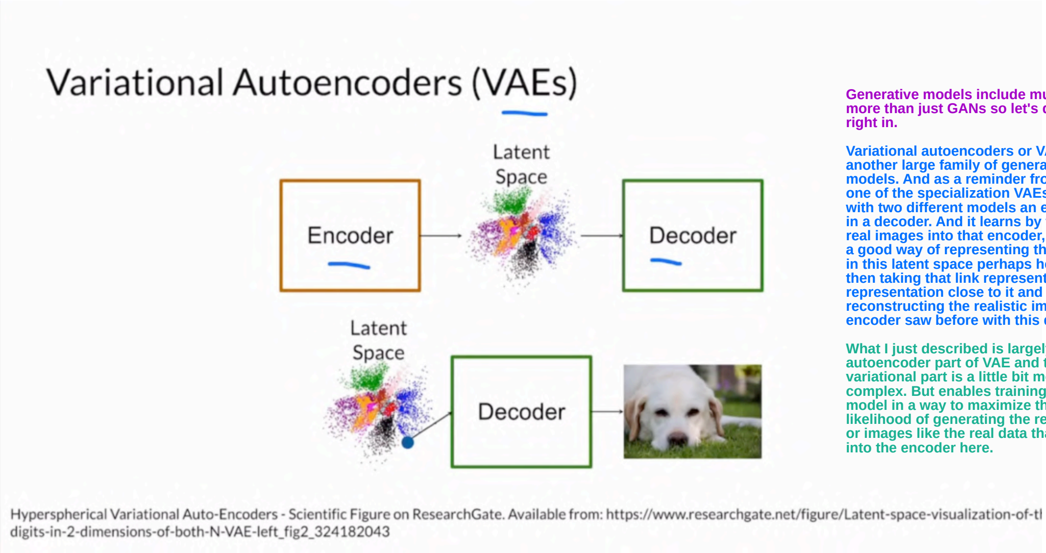
Generative Models

Noise Class Features

$$\xi, Y \to X$$

$$P(X|Y)$$

A generative model can be any machine learning model that tries to model this P of X given Y. Or if it's really just modeling that one class it's probably P of X of that data, and often it'll take in some kind of noise for that stochastic city so that you don't generate the same thing each time. And that just means variation in its outputs and a class too, then output features or objects that represent that class X.



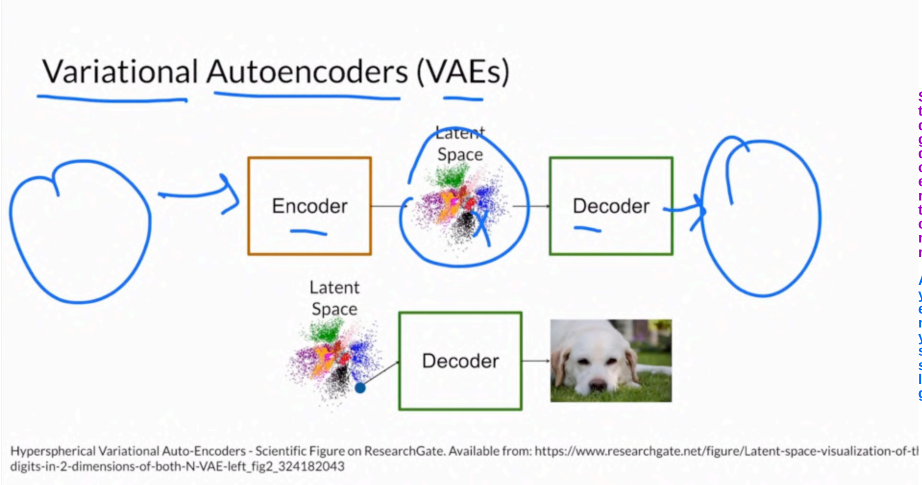
Variational Autoencoders (VAEs)

Hyperspherical Variational Auto-Encoders - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Latent-space-visualization-of-tl digits-in-2-dimensions-of-both-N-VAE-left_fig2_324182043

Generative models include much more than just GANs so let's dive right in.

Variational autoencoders or VAEs is another large family of generative models. And as a reminder from week one of the specialization VAEs work with two different models an encoder in a decoder. And it learns by feeding real images into that encoder, finding a good way of representing that image in this latent space perhaps here. And then taking that link representation or representation close to it and reconstructing the realistic image the encoder saw before with this decoder.

What I just described is largely the autoencoder part of VAE and the variational part is a little bit more complex. But enables training the model in a way to maximize the likelihood of generating the real data or images like the real data that go into the encoder here.

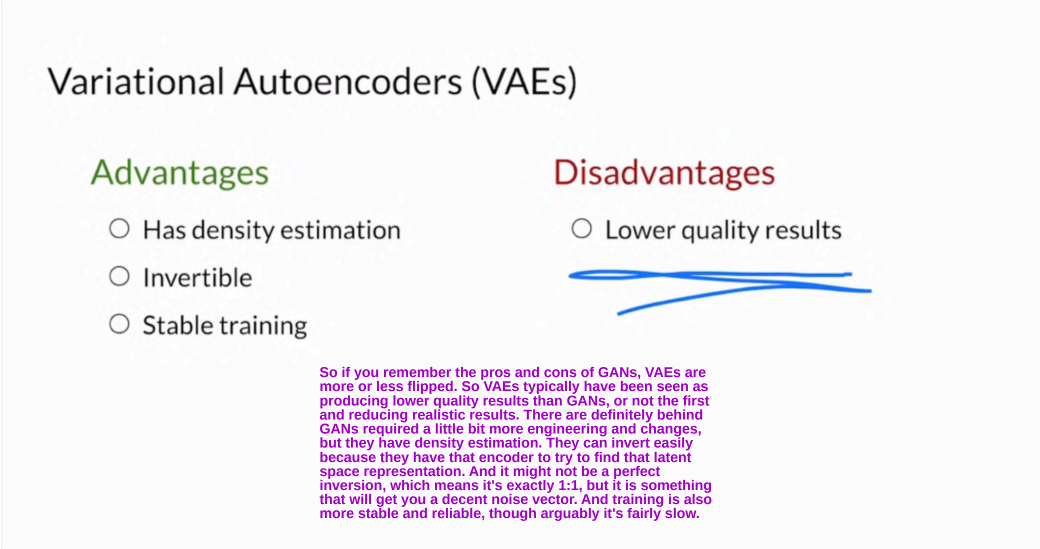


Variational Autoencoders (VAEs)

Hyperspherical Variational Auto-Encoders - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Latent-space-visualization-of-tl digits-in-2-dimensions-of-both-N-VAE-left_fig2_324182043

So at a high level VAEs is try to minimize the divergent between the generated and the real distributions. And this is often regarded as a slightly easier optimization resulting in stable or training, and this is also contended a blurrier results or lower fidelity result.

After if you train the VAE you actually loop off the encoder just like you don't need the discriminator and you use the decoder similar to the generator you sample points from your latent space in your able to generate an output image.



Variational Autoencoders (VAEs)

Advantages

○ Has density estimation

○ Invertible

○ Stable training

Disadvantages

○ Lower quality results

So if you remember the pros and cons of GANs, VAEs are more or less flipped. So VAEs typically have been seen as producing lower quality results than GANs, or not the first and reducing realistic results. There are definitely behind GANs required a little bit more engineering and changes, but they have density estimation. They can invert easily because they have that encoder to try to find that latent space representation. And it might not be a perfect inversion, which means it's exactly 1:1, but it is something that will get you a decent noise vector. And training is also more stable and reliable, though arguably it's fairly slow.

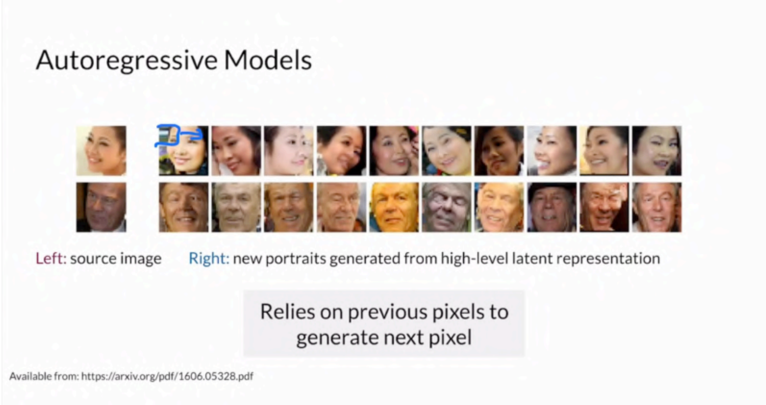
Variational Autoencoders (VAEs)

Available from: https://arxiv.org/abs/1906.00446

But the GANs camp will say, well, all that is great, but that's no use if you can't generate good samples, which is here, done, done, done.

As a result, a lot of work has been put in to make their results better, and so here's an example of a very recent VAE called VQ-VAE2 on the left and BigGAN on the right. And you can see the quality of BigGAN is slightly higher, but VAEs are beginning to have better results as well, particularly in diversity here, as you see in this generated fish.

Also this VAE esque model the VQ-VAE2 borrows many concepts in VAES, but it actually isn't considered a pure VAE solution. In fact, it relies on an autoregressive network component too



Autoregressive Models

Left: source image Right: new portraits generated from high-level latent representation

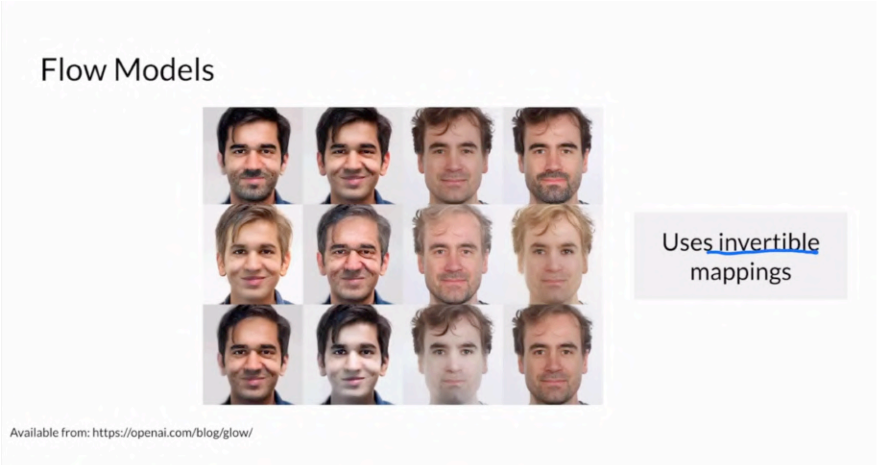
Relies on previous pixels to generate next pixel

Available from: https://arxiv.org/pdf/1606.05328.pdf

What an autoregressive model is. It's a model that looks like previous pixels to determine the next pixel. So maybe it sees a few pixels here and then it's able to determine the rest of the pixels for that image. And this is another type of generative model, and it goes pixel by pixel based on the previous pixel, and so you can think of it as it's conditioning on the previous pixels or what's the next pixel.

It can't see into future pixels, so it can't see into future pixels it can only look at past pixels. If you're familiar with RNNs or language and speech models, it's a very similar to that concept as well, where you can see into the future.

As you can probably tell this model is not fully unsupervised because it depends on those previous pixels. So it is a supervised technique, meaning it will require anchor pixels to start generating, can't generate from noise.



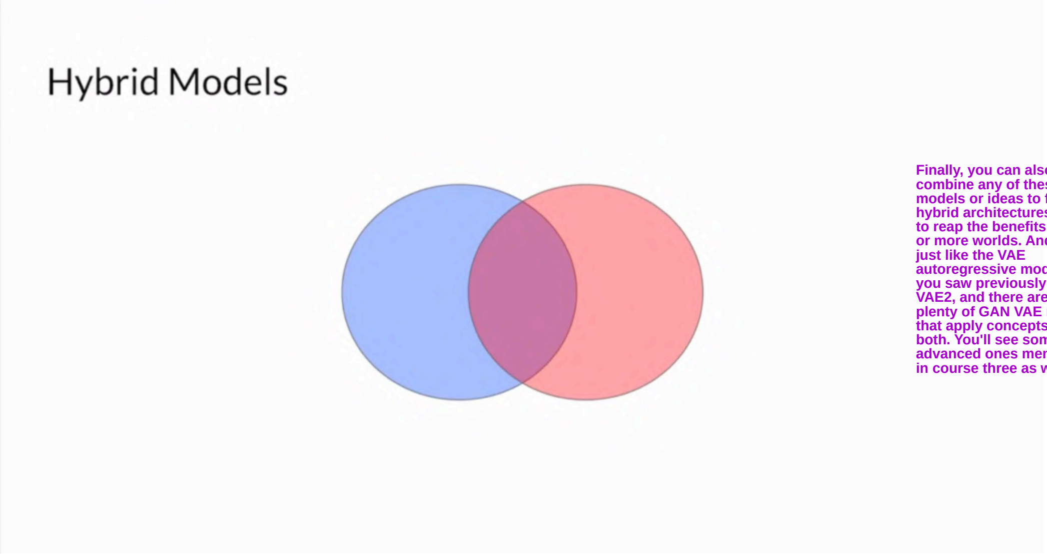
Flow Models

Uses invertible mappings

Available from: https://openai.com/blog/glow/

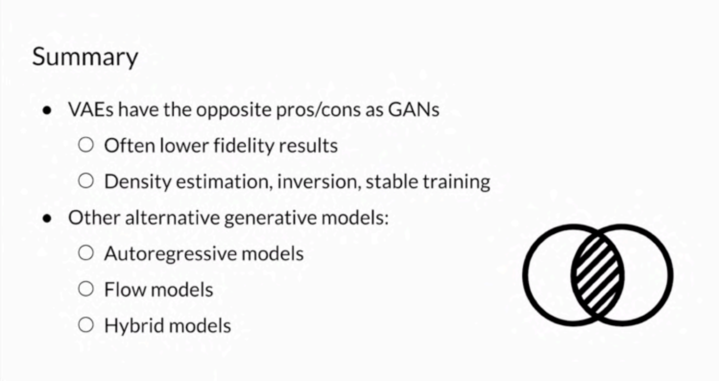
Another type of generative model is a flow model, and these are hard and long to train. But it's been a very cool new idea that's based on likelihood defined in invertible mapping between the noise in the generated image. And so obviously it will be invertible and at a high level what it's doing is from an initial simple distribution it finds sequences of invertible transformations to create more complex distributions.

So assume that this is starts with something very simple. It had these invertible mappings that are represented by these arrows. It gets more complex distributions and ultimately it's able to model faces, and this is an example flow model called glow.



Hybrid Models

Finally, you can also combine any of these models or ideas to form hybrid architectures to try to reap the benefits of two or more worlds. And that's just like the VAE autoregressive model that you saw previously called VAE2, and there are also plenty of GAN VAE models that apply concepts from both. You'll see some advanced ones mentioned in course three as well.



Summary

- VAEs have the opposite pros/cons as GANs
  - ○ Often lower fidelity results
  - ○ Density estimation, inversion, stable training
- Other alternative generative models:
  - ○ Autoregressive models
  - ○ Flow models
  - ○ Hybrid models

So in summary, VAEs have more or less the opposite pros/cons list as GANs. Notably, the results are generally blurrier, though that's arguable, but it can estimate density, invert easily, and has stable training.

However, GANs have improved on these disadvantages in many ways as training a stabilized greatly and approximate inversion, which is what you need for editing and image has been reduced to an engineering problem of finding your Z vector through another model.

VAEs have also come along way to getting better results, so all in all when it comes to applications, I'd say GANs are still more useful when realistic generation is the main goal.

As you saw in this section, other alternative generative models include the autoregressive model, flow models and also hybrid models of all of these.

Machine Bias

Before going into the discussion on bias in machine learning, please read this case study to gain an understanding of the impact these biases can have on real lives:

Machine Bias (Angwin, Larson, Mattu, and Kirchner, 2016): https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing deeplearning.ai

Intro to Machine Bias

In the following section, I'll take a step back from GANs and discuss bias. An issue that's seen around the world and seeps into many aspects of life to which machine learning and GANs are no exception. The purpose of these sections is to bring awareness to bias in machine learning and more specifically in GANs, which is the first step towards eliminating bias in your models.

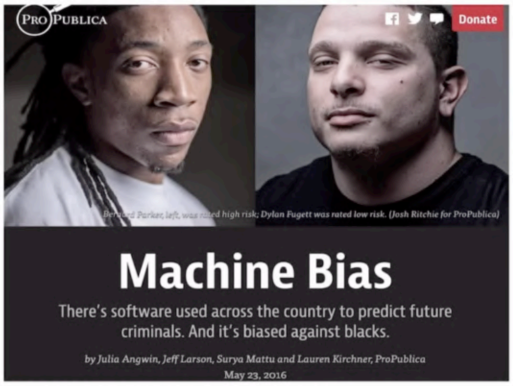
Outline

- *Machine Bias* (ProPublica)
- Racial disparity in AI for risk assessments
- Impacts of biased AI

You get a brief introduction to the machine bias article in the journal ProPublica, and a discussion of how racial disparity was found in proprietary software that's used across the country in criminal sentencing and the impact that this has.

Machine Bias

Risk assessment = likelihood of committing a crime in the future



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

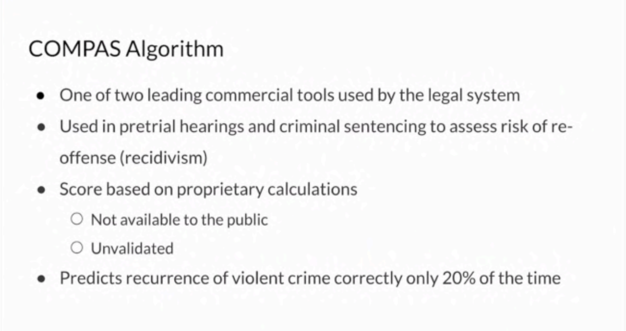
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

First off, I'll go over the main points from the ProPublica news article called Machine Bias. In the United States criminal justice system, courts are increasingly relying on risk assessments or the likelihood you're going to commit a crime in the future. Many risk score calculations are now computerized and they're starting to rely on machine learning more and more.

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



COMPAS Algorithm

- One of two leading commercial tools used by the legal system
- Used in pretrial hearings and criminal sentencing to assess risk of re-offense (recidivism)
- Score based on proprietary calculations
  - ○ Not available to the public
  - ○ Unvalidated
- Predicts recurrence of violent crime correctly only 20% of the time

ProPublica used a public records request to assess one of the two leading commercial models for it's machine bias series, and this algorithm is called COMPAS.

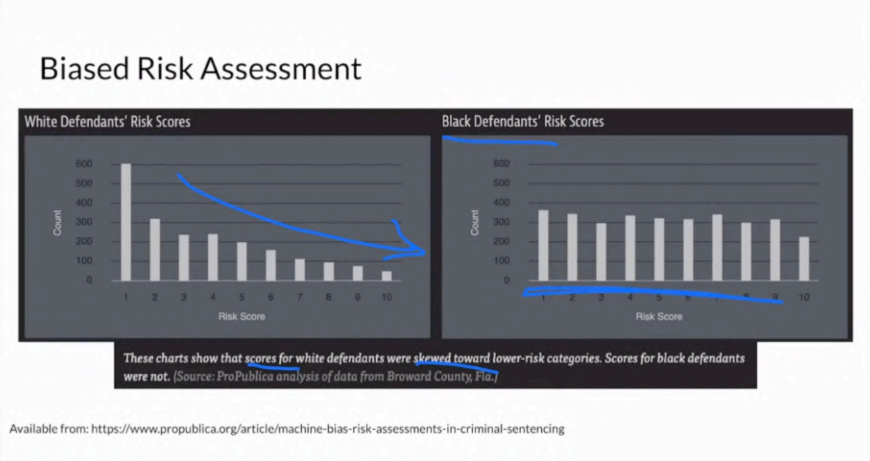
The purpose of COMPAS was to make pretrial sentencing less biased, so its purpose was to reduce bias, but ProPublica's findings suggest that the algorithm itself is actually significantly biased.

One important thing to note about COMPAS is that the calculations used for the scores are considered proprietary and the company doesn't share exactly what they are. The calculations are broadly based on a questionnaire filled out by defendants, as well as on their criminal records. There's no question about race there, but it asks questions like, was one of your parents ever sent to jail, and is it wrong for a hungry person to steal? So getting a little bit philosophical and maybe moral.

Then judges can then increase or decrease that sentence length based on their evaluation of the risk score, and so yes, these questions do not question race at all, but we'll see how some of these proxy questions might be trying to get at that or might be biased due to the type of question asked.

Unfortunately, the COMPAS algorithm only predicted recurrence of violent crime correctly about 20 percent of the time.

Biased Risk Assessment



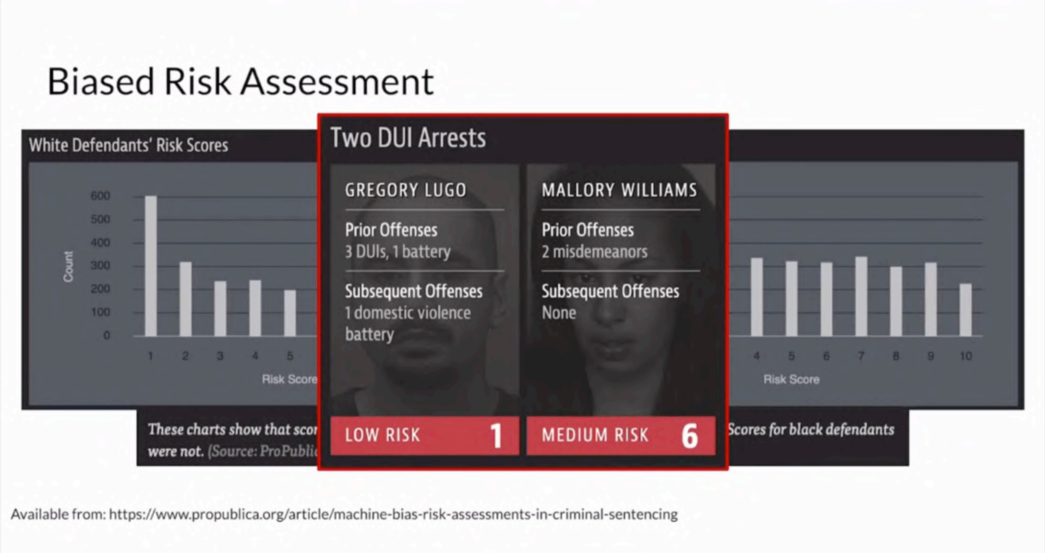
These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.)

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

ProPublica found that models predicted black defendants were 77 percent more likely and more at risk for committing a future violent crime and 45 percent more likely for any kind of future crime. You can see in the graph that while the count of white defendants in high risk scores decreased together, this graph for black defendants shows that the likelihood remains the same.

This starts to suggest that the scores for white defendants were skewed towards these lower risk categories while the scores for black defendants were not.

Biased Risk Assessment



Two DUI Arrests

GREGORY LUGO

Prior Offenses
3 DUIs, 1 battery

Subsequent Offenses
1 domestic violence battery

LOW RISK 1

MALLORY WILLIAMS

Prior Offenses
2 misdemeanors

Subsequent Offenses
None

MEDIUM RISK 6

These charts show that scor ... *Scores for black defendants*
were not. (Source: ProPubli

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Consequences of a Higher Score

Paul Zilly

Plea deal overturned and sentenced to two years in state prison.

"Had I not had the COMPAS, I believe it would likely be that ***I would have given one year, six months***"

- Appeals judge

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Consequences of a Higher Score

Sade Jones

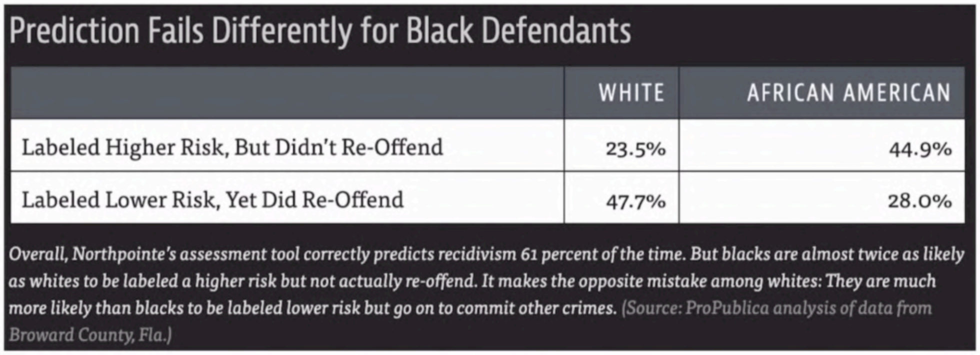
Bond was raised from the recommended $0 to $1000

"I went to McDonald's and a dollar store, and they all said no ***because of my background***"

- Jones

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Prediction Failure

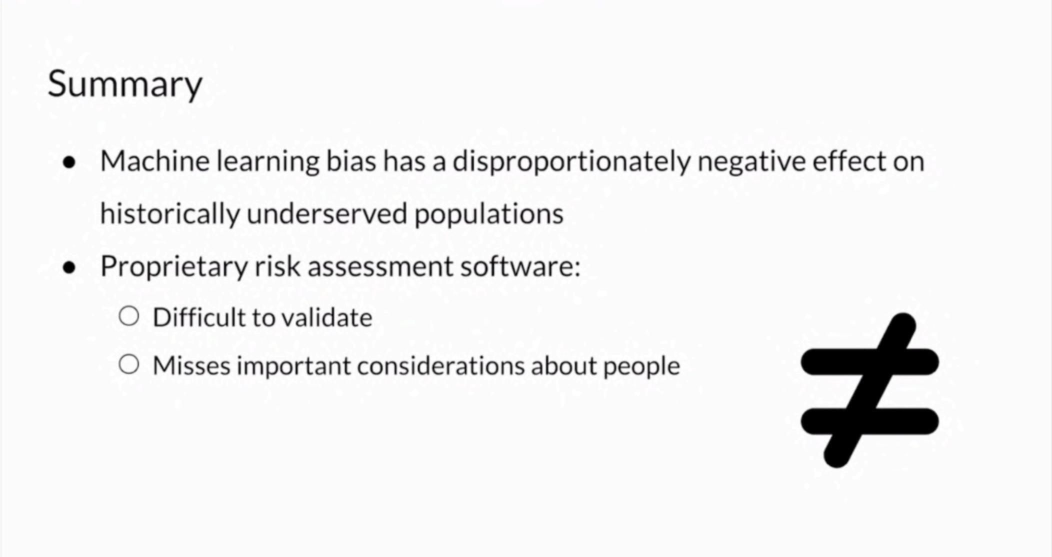


Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



Summary

- Machine learning bias has a disproportionately negative effect on historically underserved populations
- Proprietary risk assessment software:
    - ○ Difficult to validate
    - ○ Misses important considerations about people

≠

Fairness Definitions

To understand some of the existing definitions of fairness and their relationships, please read the following paper and view the Google glossary entry for fairness:

1. Fairness Definitions Explained (Verma and Rubin, 2018): https://fairware.cs.umass.edu/papers/Verma.pdf
2. Machine Learning Glossary: Fairness (2020): https://developers.google.com/machine-learning/glossary/fairness

A Survey on Bias and Fairness in Machine Learning

To understand the complex nature of bias and fairness, please read the following paper describing ways they exist in artificial intelligence and machine learning:

A Survey on Bias and Fairness in Machine Learning (Mehrabi, Morstatter, Saxena, Lerman, and Galstyan, 2019): https://arxiv.org/abs/1908.09635

deeplearning.ai

Defining Fairness

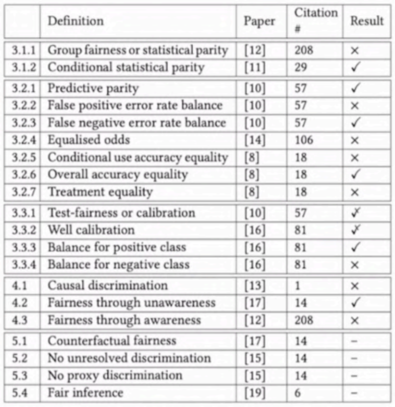
Outline

- What is fairness?
- Complexity of defining fairness

Fairness in Machine Learning

Reading 1: Fairness Definitions Explained

Reading 2: A Survey on Bias and Fairness in Machine Learning

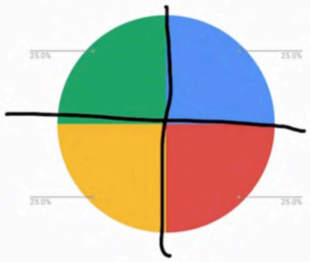


	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

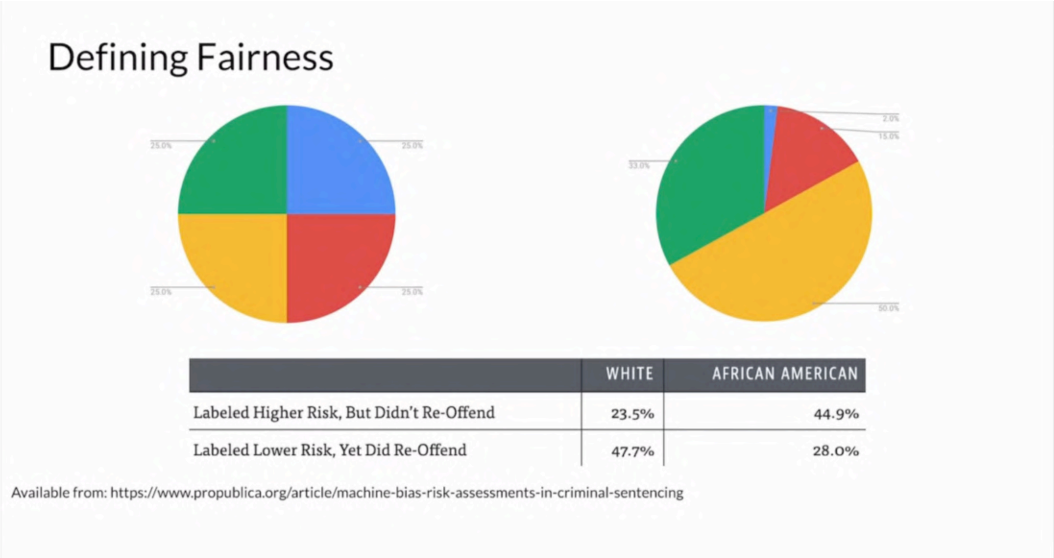
Table 1: Considered Definitions of Fairness

Available from: https://fairware.cs.umass.edu/papers/Verma.pdf

Defining Fairness



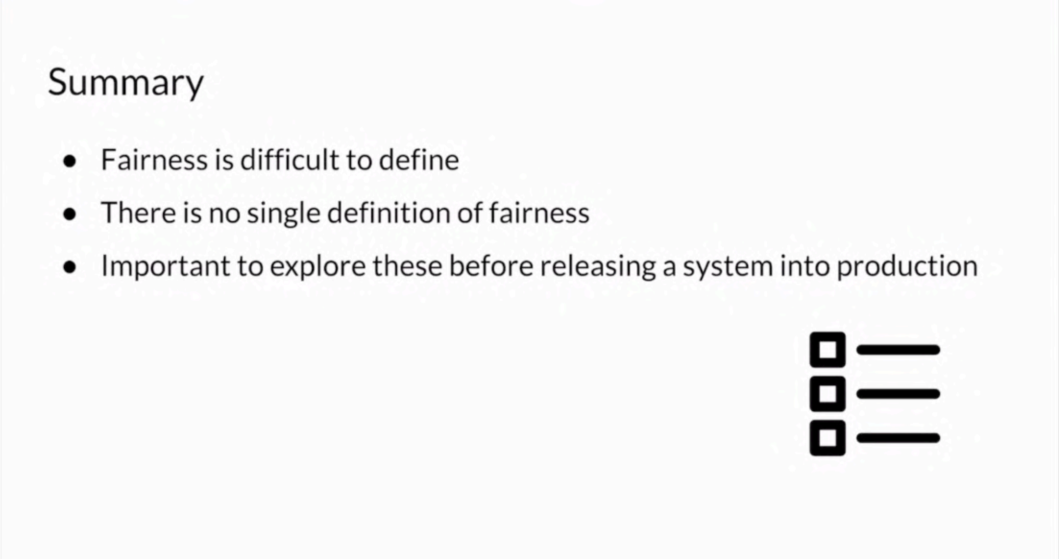
demographic parity



Defining Fairness

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



Summary

- Fairness is difficult to define
- There is no single definition of fairness
- Important to explore these before releasing a system into production

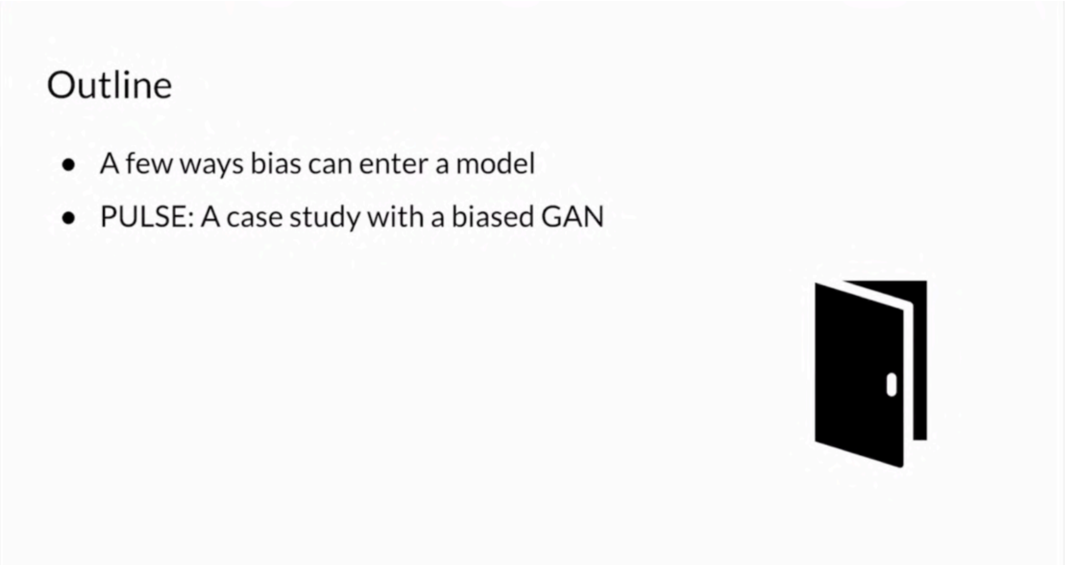
Finding Bias

Now that you've seen how complex fairness is, how do you find bias in existing material (models, datasets, frameworks, etc.) and how can you prevent it? These two readings offer some insight into how bias was detected and some avenues where it may have been introduced.

1. Does Object Recognition Work for Everyone? (DeVries, Misra, Wang, and van der Maaten, 2019): https://arxiv.org/abs/1906.02659
2. What a machine learning tool that turns Obama white can (and can't) tell us about AI bias (Vincent, 2020): https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

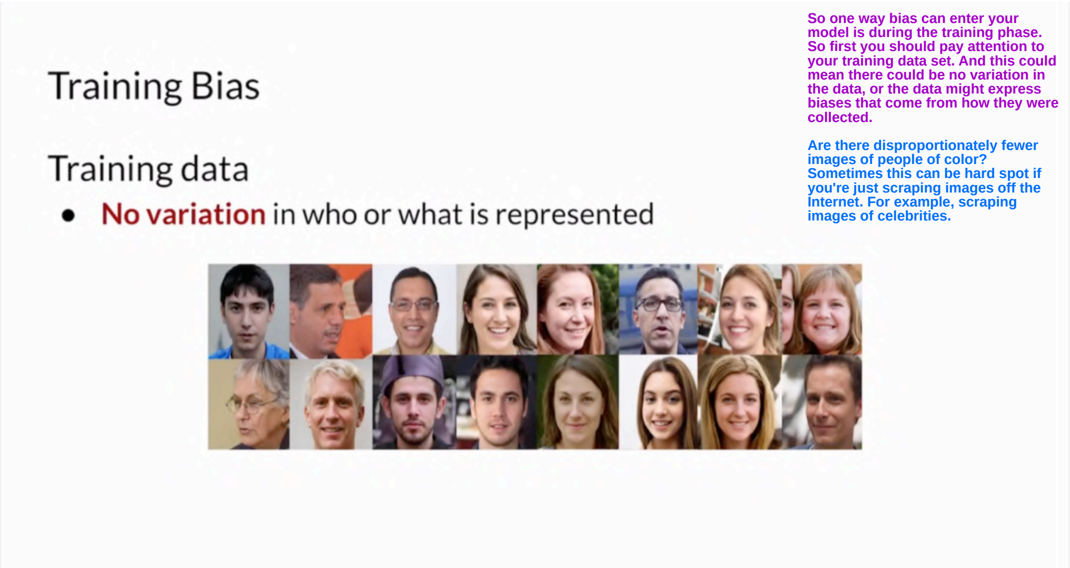
deeplearning.ai

Ways Bias is Introduced



Outline

- A few ways bias can enter a model
- PULSE: A case study with a biased GAN



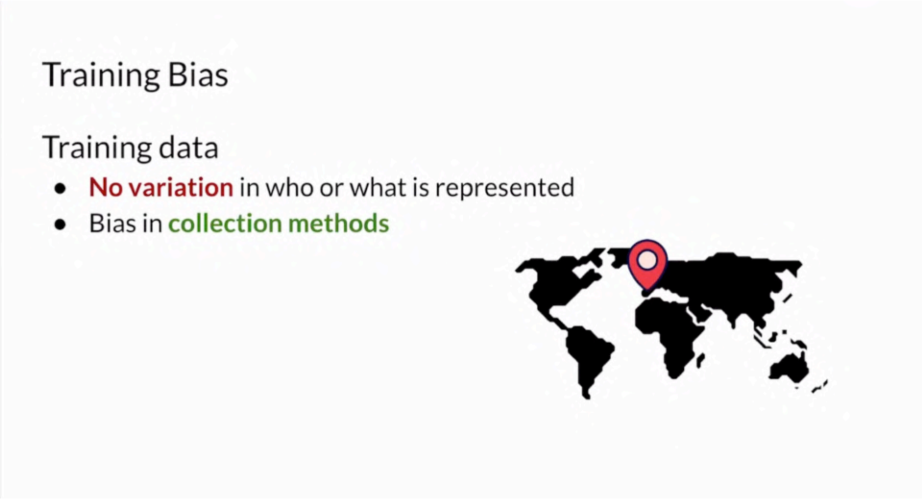
Training Bias

Training data

- **No variation** in who or what is represented

So one way bias can enter your model is during the training phase. So first you should pay attention to your training data set. And this could mean there could be no variation in the data, or the data might express biases that come from how they were collected.

Are there disproportionately fewer images of people of color? Sometimes this can be hard spot if you're just scraping images off the Internet. For example, scraping images of celebrities.

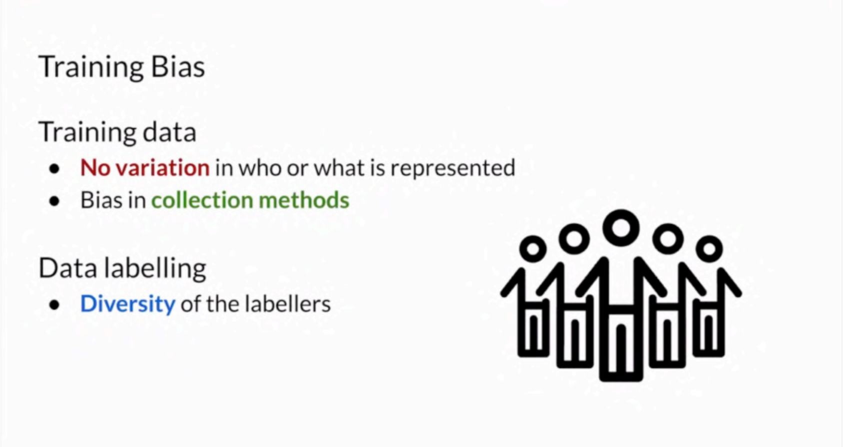


Training Bias

Training data

- **No variation** in who or what is represented
- Bias in **collection methods**

You also need to consider how your data was collected, was the data all collected from one location, one web scrape? Was it collected by a single person, or a single demographic of people. Remember that whatever is considered a quote, unquote diverse data set. Also really depends on your definition of fairness



Training Bias

Training data

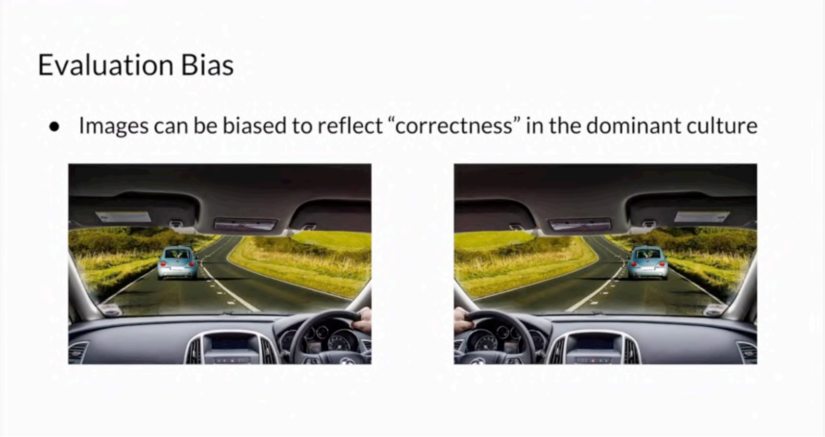
- **No variation** in who or what is represented
- Bias in **collection methods**

Data labelling

- **Diversity** of the labellers

If you're using label data, then the diversity of the labellers impact your data as well. And this is because different demographics might label things differently, and that might cause inherent biases to arise in your data.

For example, when labellers are mostly men, labeling resumes of people as worthy of an interview for a software engineering rule or not. These are just a few considerations fusion make when you're preparing your models training data.



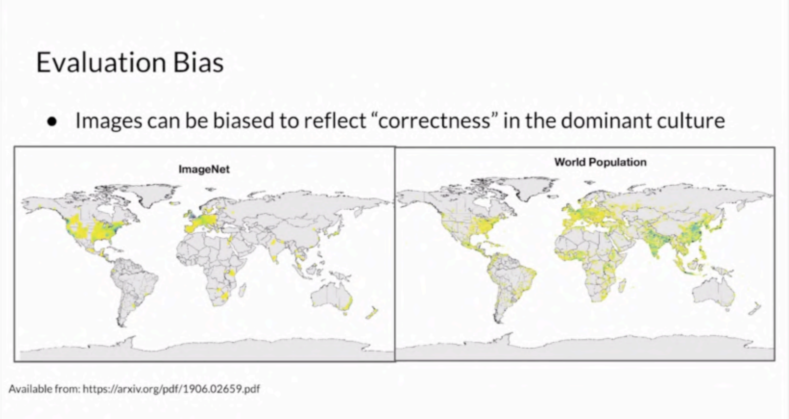
Evaluation Bias

- Images can be biased to reflect “correctness” in the dominant culture

Biases that exist in broader society can also shape all portions of your model. In other area it can be introduced is during your evaluation, such as who created the evaluation method you're using.

Evaluation methods could be biased towards images that are often regarded as quote, Unquote. Good or correct in that society or in one culture, but not another.

A simple example is whether cars are driven on the left or right side of the road. If they are commonly driven on a certain side, then the evaluation may reflect that and poorly evaluate or score ones. That have the wheel on the quote, unquote wrong side. This would solely be dependent on local driving laws.



Evaluation Bias

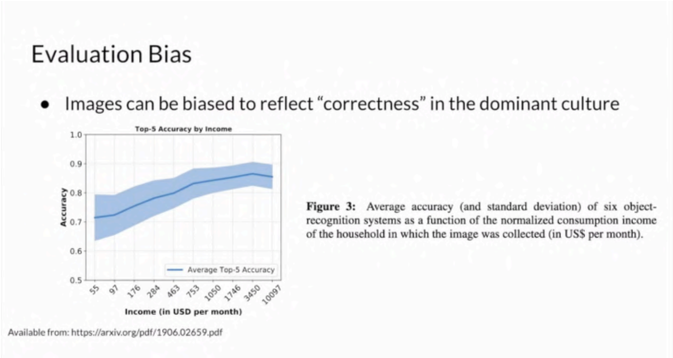
- Images can be biased to reflect "correctness" in the dominant culture

Available from: https://arxiv.org/pdf/1906.02659.pdf

Another more concrete example, pertains to image net. The data set used to train the inception V3 model, which FID uses to evaluate Gantz. As well as other evaluation metrics.

More than half of the images in ImageNet, come from the USA and Great Britain. Compared to the population densities of the world, this is not really representative of where most people are from. And that imbalance leads systems to inaccurately classify images into categories that differ by geography. Would arise had be classified as hair, or a poncho as a scarf.

The way evaluation calculations are computed, can reinforce biases within the model you develop. And can make you think a model is great at a certain task, when it actually is unable to perform that task.



Evaluation Bias

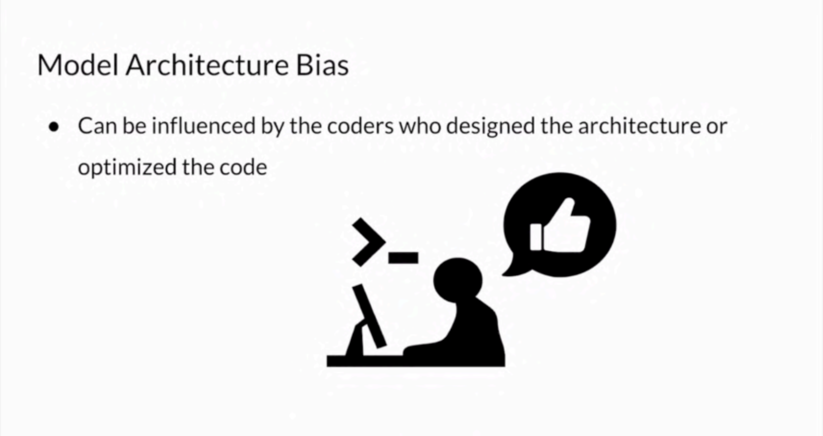
- Images can be biased to reflect "correctness" in the dominant culture

Figure 3: Average accuracy (and standard deviation) of six object-recognition systems as a function of the normalized consumption income of the household in which the image was collected (in US$ per month).

Available from: https://arxiv.org/pdf/1906.02659.pdf

For example, researchers actually took items from inside household that had fairing income levels. Then they evaluated the accuracy of the top object recognition models, on these products. The result was that the images taken from higher income families, had higher accuracy on these models.

So what's not great about this, is that the people who developed these models might have concluded. That they had great models for seeing the world. Visual perception that is human level. And saying that visual perception is solved, when really it's only reached a high accuracy level for objects of a particular socioeconomic status. And that's not the perception I would envision for a great model.

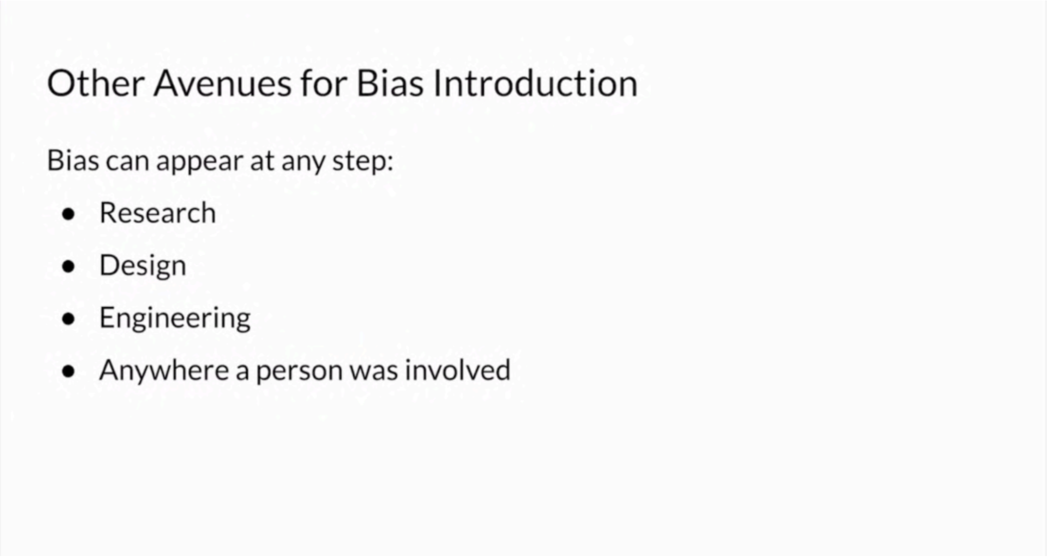


Model Architecture Bias

- Can be influenced by the coders who designed the architecture or optimized the code

So now, bias can be introduced through the architecture as well. What was the diversity of programmers who optimize the code? What were their views on what is quote, unquote right, or quote, unquote wrong? And what looks good or bad can impact the images generated? After all especially in generative models, where the evaluation metrics aren't great. It's even more important to lend a critical eye, to how various problems are chosen in this field. Because once you choose important problems, people will optimize solutions to those definitions of right or wrong, of good or bad. And that will angle certain directions of research, and how their chosen as well.

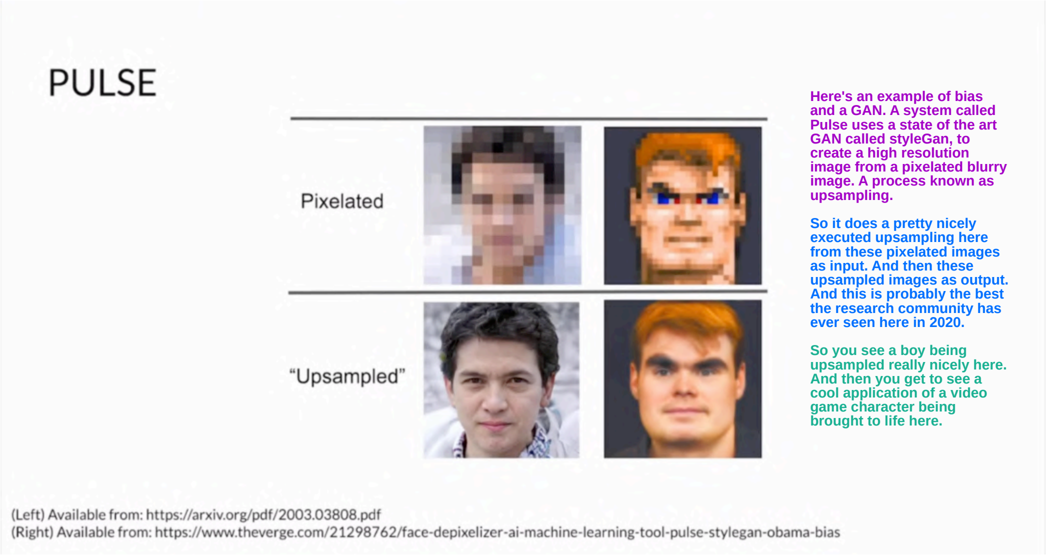
So as an example, the loss function used can skew what a model thinks is correct, for GAN generating faces. This could be the difference between having a more light colored or more dark colored skin.



Other Avenues for Bias Introduction

Bias can appear at any step:

- Research
- Design
- Engineering
- Anywhere a person was involved



PULSE

Pixelated

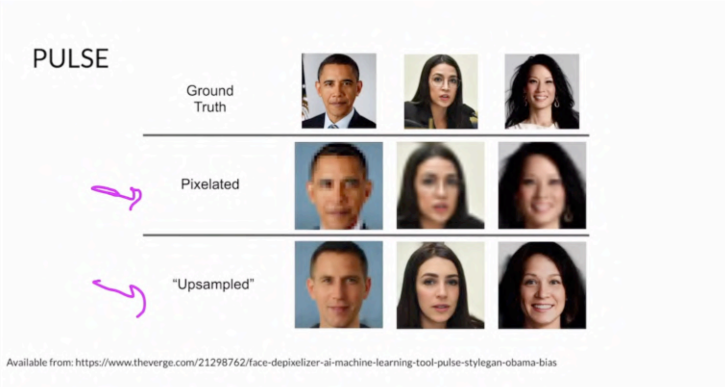
“Upsampled”

(Left) Available from: https://arxiv.org/pdf/2003.03808.pdf
(Right) Available from: https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

Here's an example of bias and a GAN. A system called Pulse uses a state of the art GAN called styleGan, to create a high resolution image from a pixelated blurry image. A process known as upsampling.

So it does a pretty nicely executed upsampling here from these pixelated images as input. And then these upsampled images as output. And this is probably the best the research community has ever seen here in 2020.

So you see a boy being upsampled really nicely here. And then you get to see a cool application of a video game character being brought to life here.



PULSE

Ground Truth

Pixelated

“Upsampled”

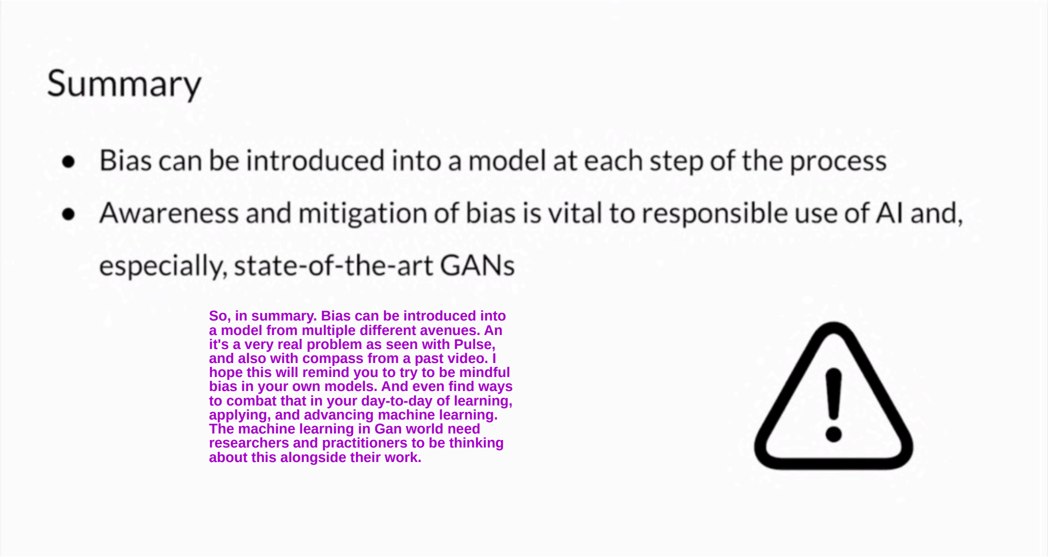
Available from: https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

However, that's not all. An example of a pixelated Obama photo who is biracial, is upsampled to a distinctly white man. In addition, politician Alexandria Ocasio Cortez, and actress Lucy Liu, are also transformed into unrecognizable versions of themselves. That are arguably more quote, unquote White in ethnicity. More closely resembling the average phase, from the style game generator in what it was trained on.

Performing better on a video game character who is White. Than these people of color is a fairly strong indication there is a problem with bias in the model. But it's hard to say where the system failed. Is it StyleGan or is it the system that was built on top of StyleGan Pulse. Or was it the data set that StyleGan was trained on?

There's been research to mitigate bias in GANs, such as using an adversarial loss to punish models for being biased. It's complicated and an important area of work. So I really encourage you to go check that out.

As of now, these types of issues are starting to become more spotlighted in the machine learning community. And the authors of the Pulse Paper, have since put out a statement of caution with applying their model.



Summary

- Bias can be introduced into a model at each step of the process
- Awareness and mitigation of bias is vital to responsible use of AI and, especially, state-of-the-art GANs

So, in summary. Bias can be introduced into a model from multiple different avenues. An it's a very real problem as seen with Pulse, and also with compass from a past video. I hope this will remind you to try to be mindful bias in your own models. And even find ways to combat that in your day-to-day of learning, applying, and advancing machine learning. The machine learning in Gan world need researchers and practitioners to be thinking about this alongside their work.

Works Cited

All of the resources cited in Course 2 Week 2, in one place. You are encouraged to explore these papers/sites if they interest you, especially because this is an important topic to understand. They are listed in the order they appear in the lessons.

From the videos:

- Hyperspherical Variational Auto-Encoders (Davidson, Falorsi, De Cao, Kipf, and Tomczak, 2018): https://arxiv.org/abs/1804.00891
- Generating Diverse High-Fidelity Images with VQ-VAE-2 (Razavi, van den Oord, and Vinyals, 2019): https://arxiv.org/abs/1906.00446
- Conditional Image Generation with PixelCNN Decoders (van den Oord et al., 2016): https://arxiv.org/abs/1606.05328
- Glow: Better Reversible Generative Models (Dhariwal and Kingma, 2018): https://openai.com/blog/glow/
- Machine Bias (Angwin, Larson, Mattu, and Kirchner, 2016): https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Fairness Definitions Explained (Verma and Rubin, 2018): https://fairware.cs.umass.edu/papers/Verma.pdf
- Does Object Recognition Work for Everyone? (DeVries, Misra, Wang, and van der Maaten, 2019): https://arxiv.org/abs/1906.02659
- PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models (Menon, Damian, Hu, Ravi, and Rudin, 2020): https://arxiv.org/abs/2003.03808
- What a machine learning tool that turns Obama white can (and can't) tell us about AI bias (Vincent, 2020): https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

From the notebook:

- Mitigating Unwanted Biases with Adversarial Learning (Zhang, Lemoine, and Mitchell, 2018): https://m-mitchell.com/papers/Adversarial_Bias_Mitigation.pdf
- Tutorial on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR 2020 (Gebru and Denton, 2020): https://sites.google.com/view/fatecv-tutorial/schedule?authuser=0
- Machine Learning Glossary: Fairness (2020): https://developers.google.com/machine-learning/glossary/fairness
- CelebFaces Attributes Dataset (CelebA): http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html