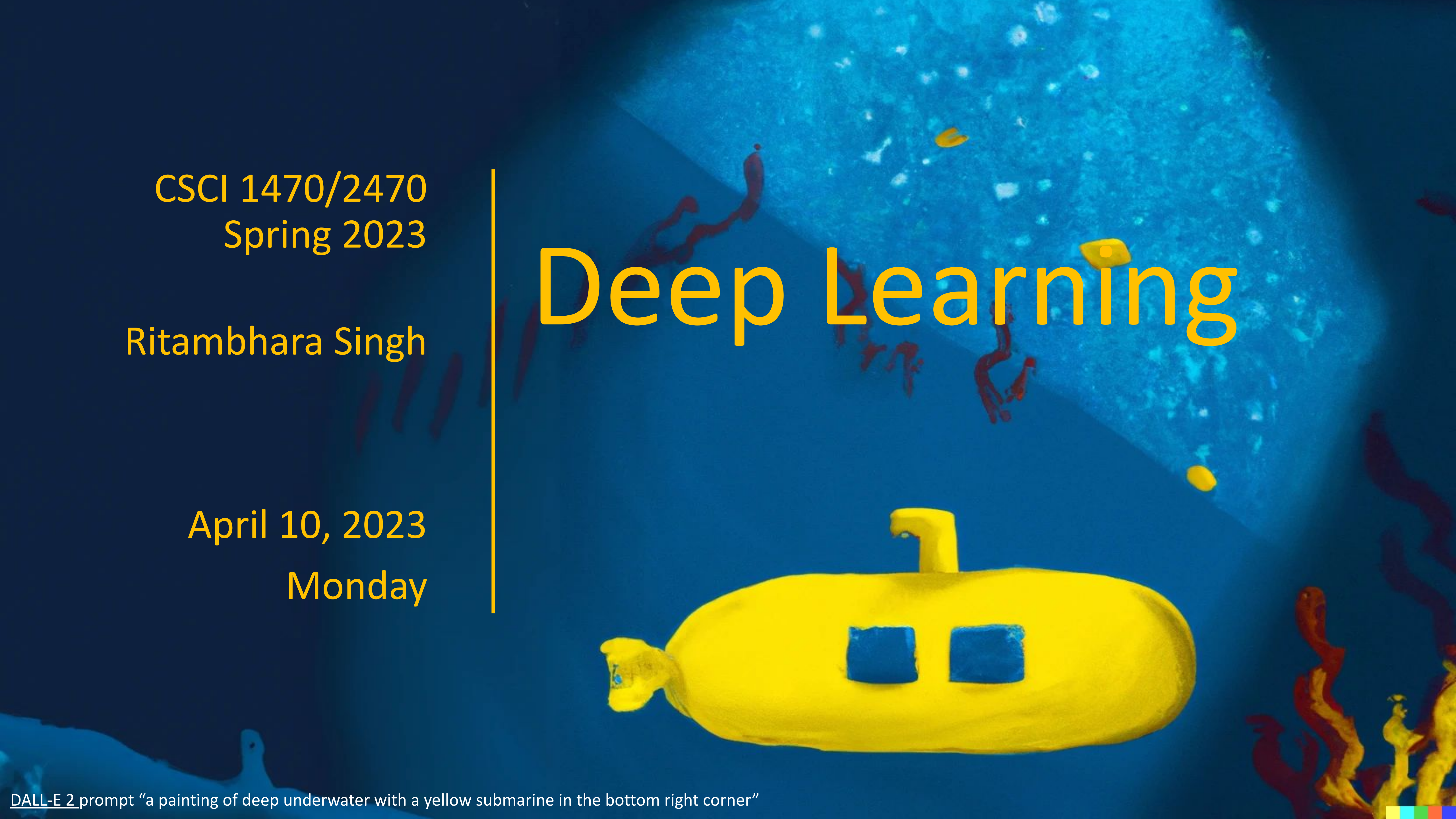


Deep Learning

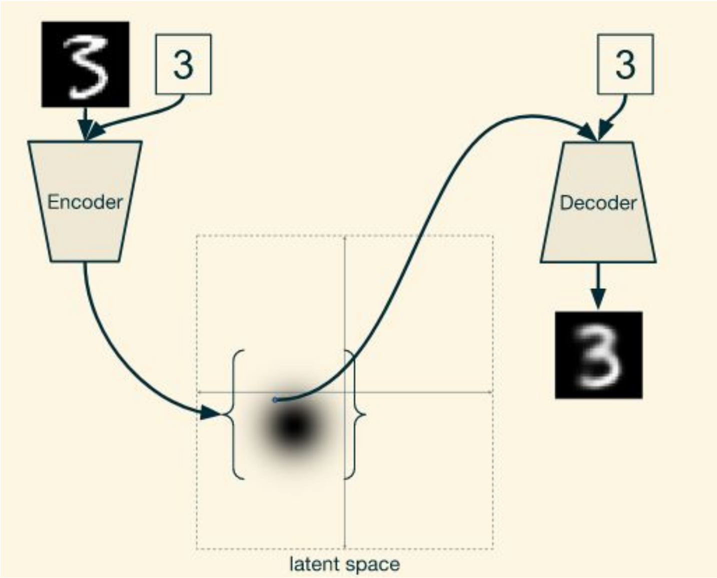
CSCI 1470/2470
Spring 2023

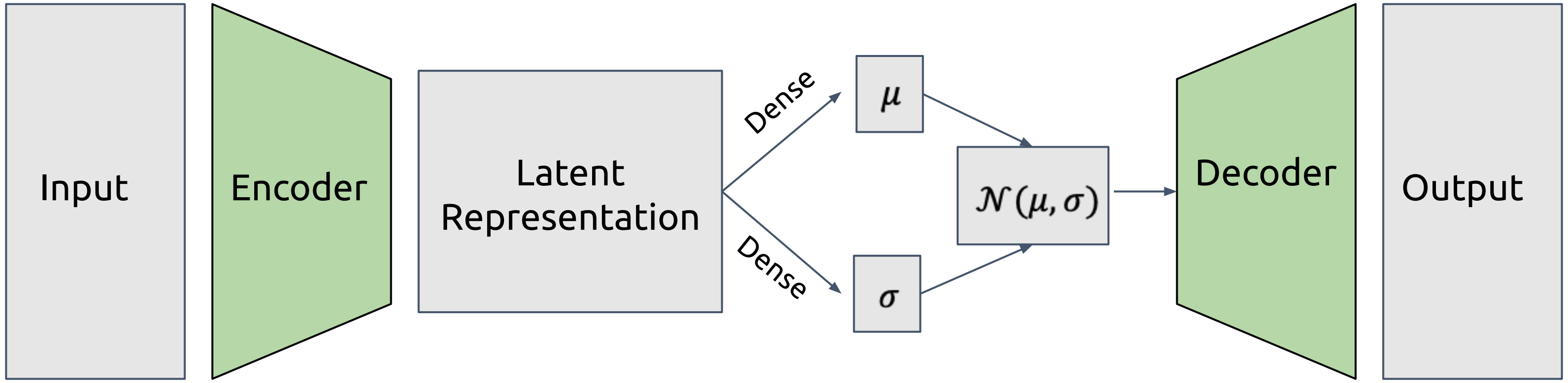
Ritambhara Singh

April 10, 2023
Monday

DALL-E 2 prompt “a painting of deep underwater with a yellow submarine in the bottom right corner”

Review: VAEs and Conditional VAEs




Input

Encoder

Latent Representation

Dense

Dense

μ

σ

$\mathcal{N}(\mu, \sigma)$

Decoder

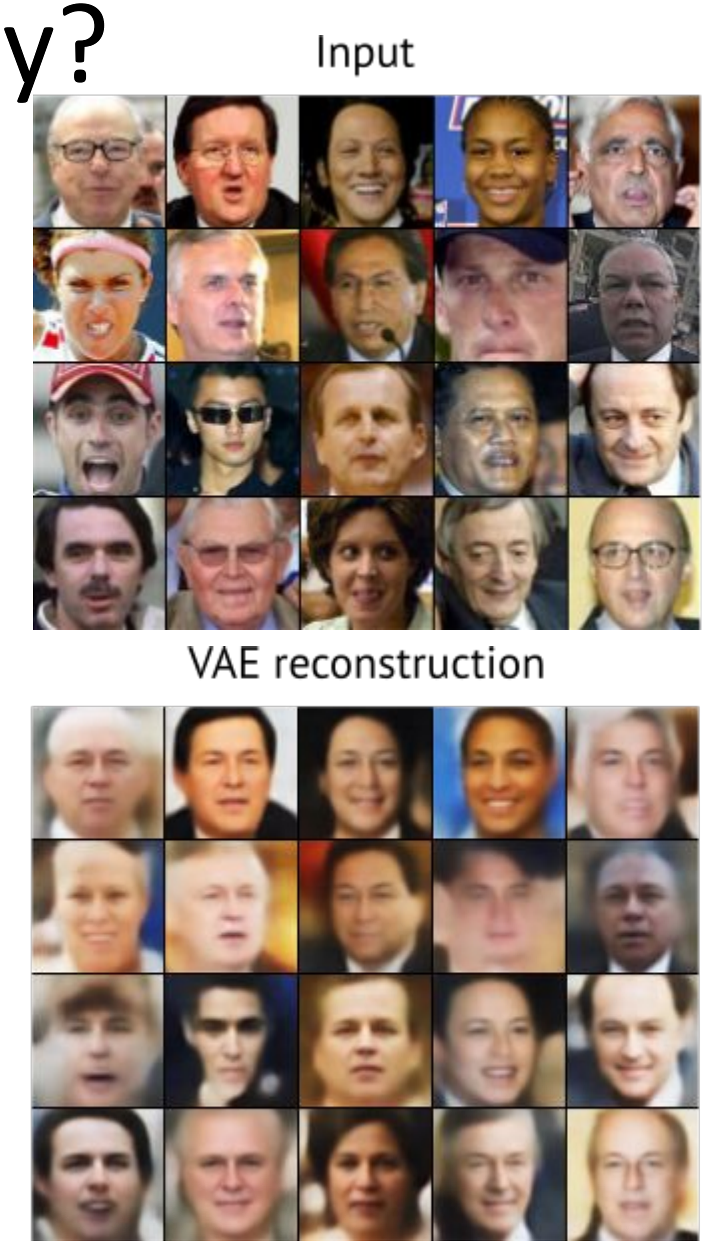
Output

https://towardsdatascience.com/understanding-conditional-variational-autoencoders-cd62b4f57bf8

Review: Why are VAE samples blurry?

- Our reconstruction loss is the culprit
- Mean Square Error (MSE) loss looks at each pixel in isolation
- If no pixel is too far from its target value, the loss won’t be too bad
- Individual pixels look OK, but larger-scale features in the image aren’t recognizable

- ***Solutions?***
  - Let’s choose a different reconstruction loss!



https://towardsdatascience.com/what-the-heck-are-vae-gans-17b86023588a

Today’s goal – learn about generative adversarial networks (GANs)

(1) Generative Adversarial Networks (GANs)

(2) Training GANs and challenges

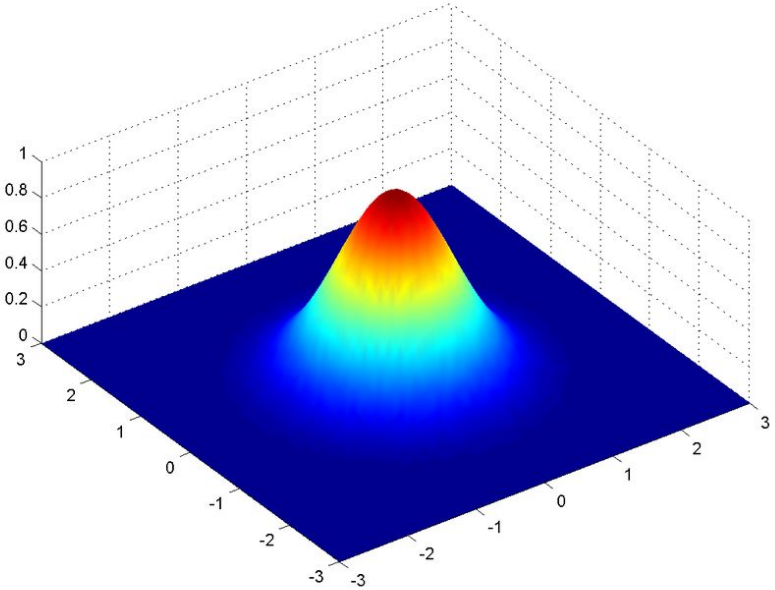
(3) Deepfakes

Generative Adversarial Networks

(a.k.a. “GANs”)

Review: A Neural Generative Model

- Input: a point $z \in \mathbb{R}^n$ drawn from the unit normal distribution $\mathcal{N}(0,1)$
- Output: a point $x \in \mathbb{R}^m$ distributed according to some more complex distribution



→ Generator Network →

The distribution of human faces

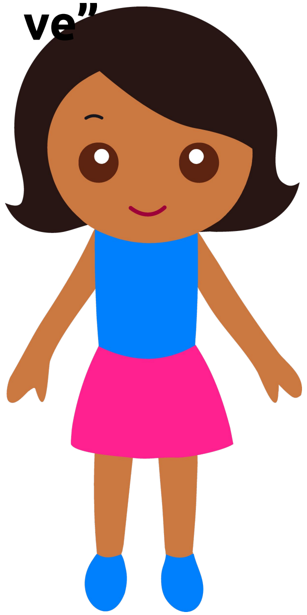
GANs by Analogy

Scenario:

Two kids are playing a detective game (”Sherlock” or “Nancy Drew”) where one of them has to fool the other in making counterfeit dollars

“Forger”

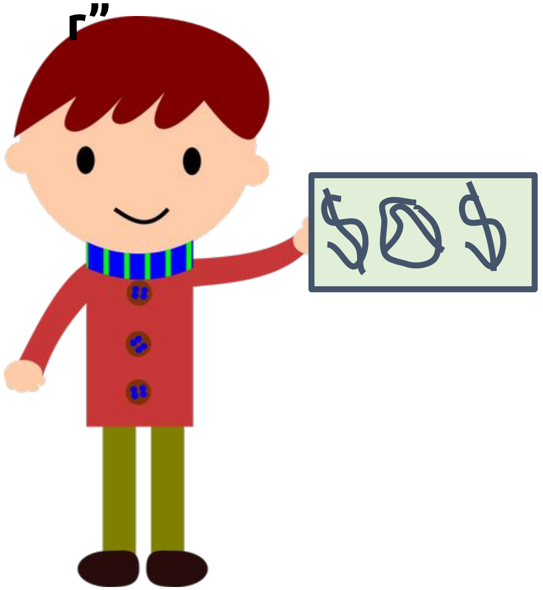
“Detective”



GANs by Analogy

- Initially, neither one of them is very good at their job
- The Forger produces horrible doodles on paper
- The Detective just looks for obvious “tells” / mistakes

“Forger”

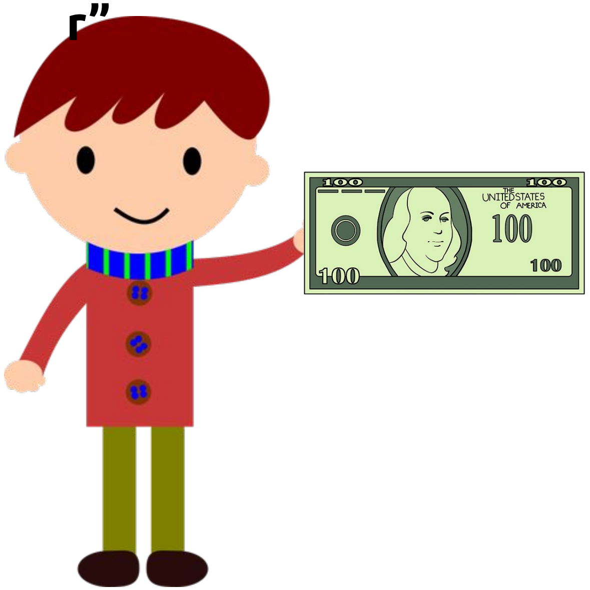


“Detective”

GANs by Analogy

- As the Detective spots the Forger’s fakes, the Forger has to devise better fakes
- The Detective, in turn, has to get better at spotting the Forger’s improved fakes

“Forger”



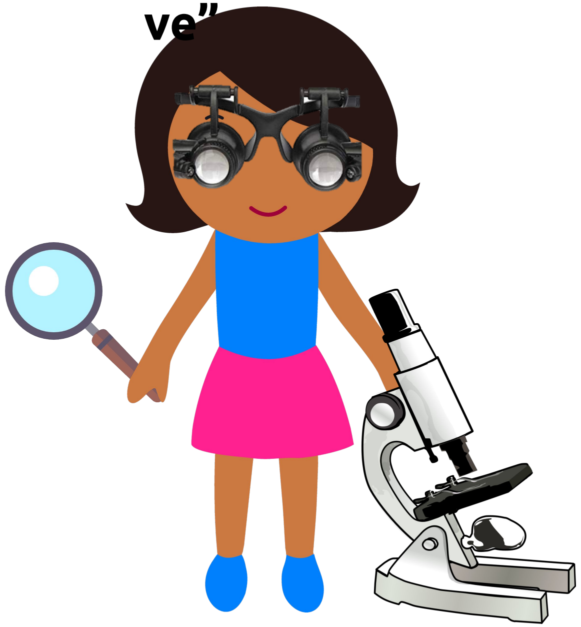
“Detective”

GANs by Analogy

- If they keep this up long enough, the Forger gets so good that their fakes are virtually indistinguishable from the real thing...
- ...and the Detective has developed ‘superhuman’ abilities to detect them

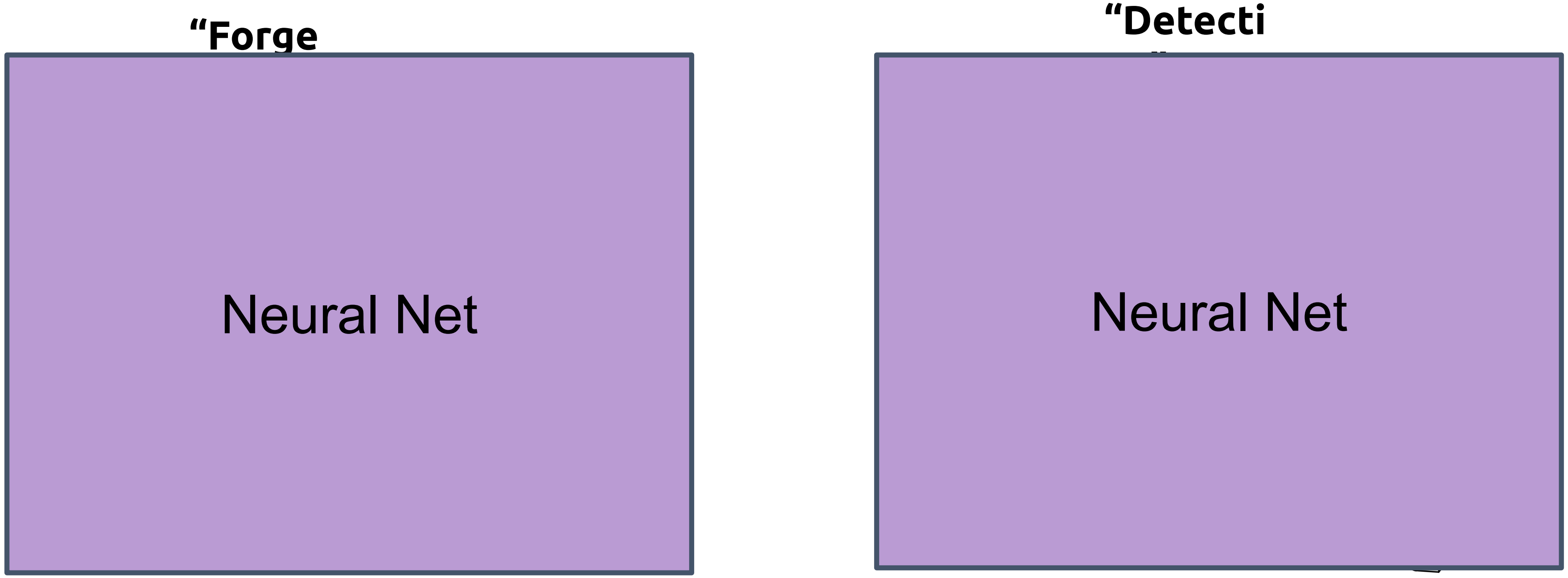
“Forger”

“Detective”



GANs by Analogy

- GANs operationalize this idea by using neural networks to serve both of these roles



“Forge

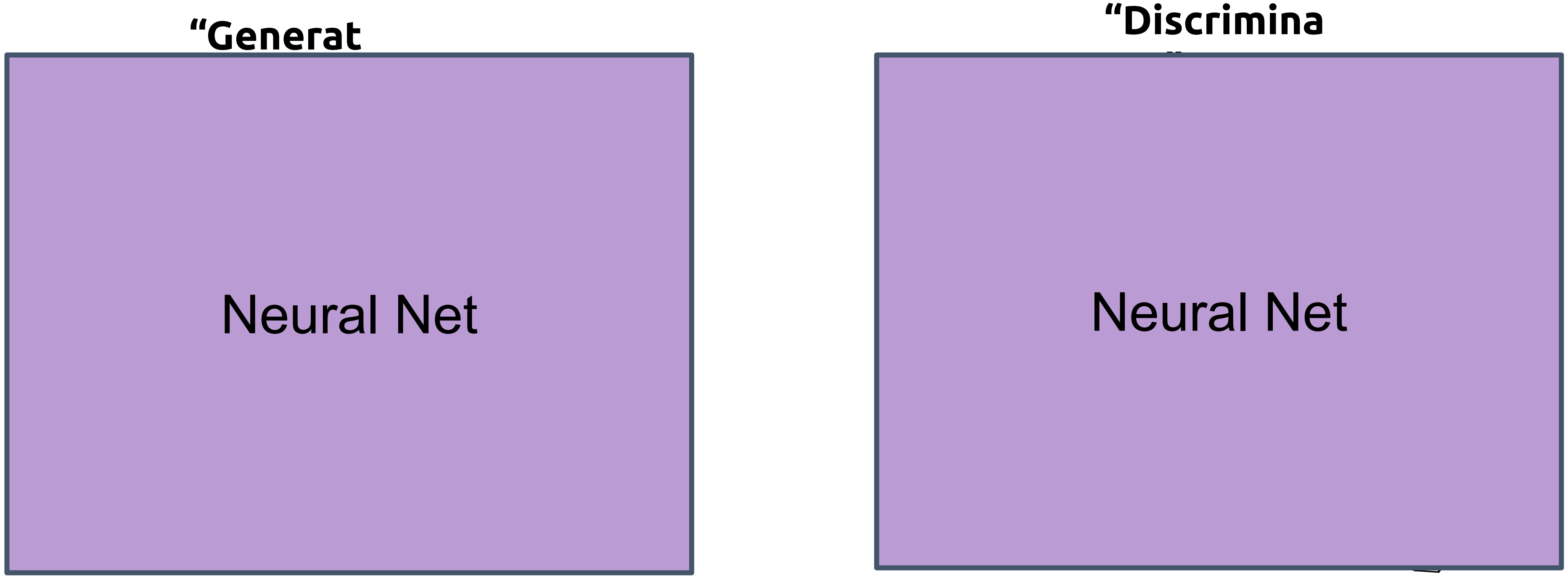
“Detecti

Neural Net

Neural Net

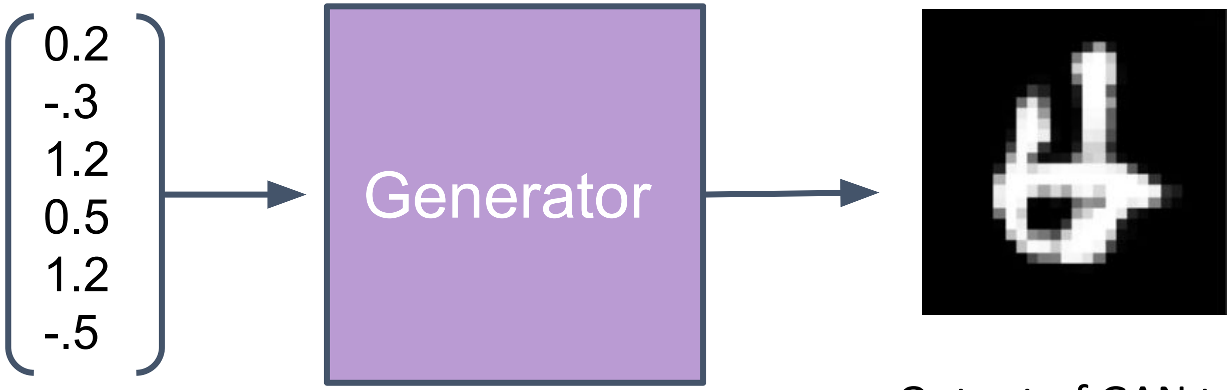
GANs by Analogy

- GANs operationalize this idea by using neural networks to serve both of these roles
- We call these networks the “Generator” and the “Discriminator”



GANs: The Generator

The generator is a neural network that takes in a random vector and produces a “fake” data point



$$\begin{pmatrix} 0.2 \\ -.3 \\ 1.2 \\ 0.5 \\ 1.2 \\ -.5 \end{pmatrix}$$

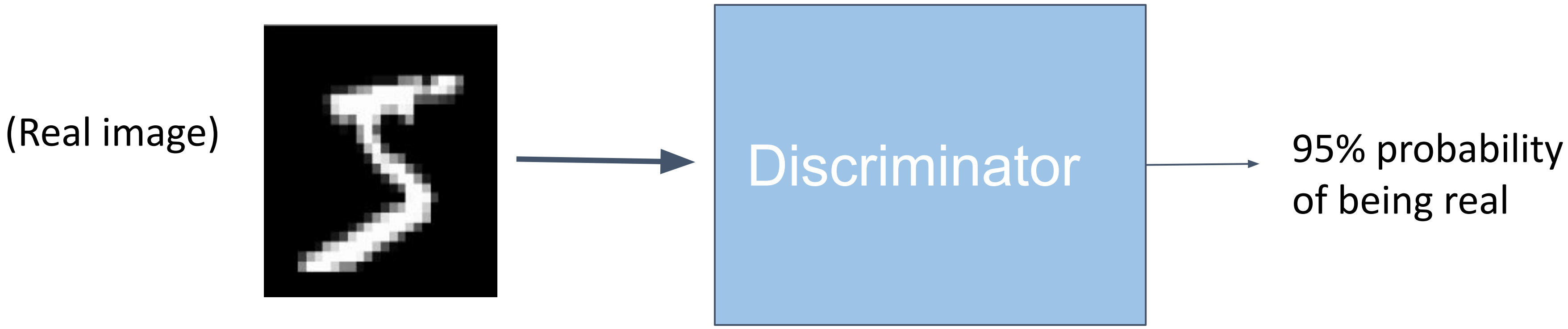
Generator

Random Vector
(sampled from unit
normal distribution)

Output of GAN trained
on MNIST images

GANs: The Discriminator

The discriminator is a neural network that takes in images and predicts the probability that the image is real:



(Real image)

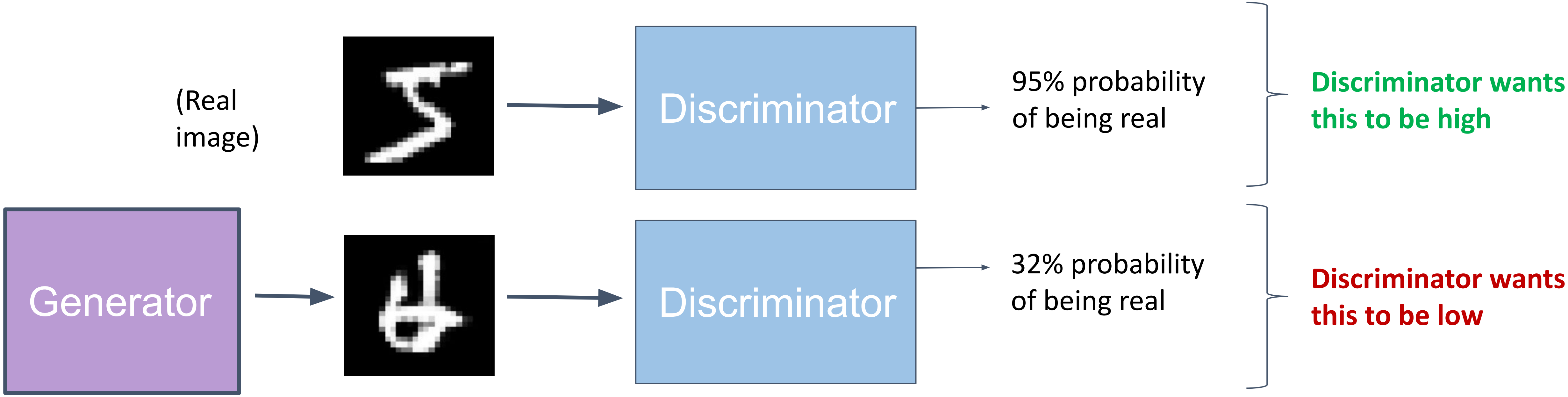
Discriminator

95% probability of being real

GANs: Training the Discriminator

Discriminator wants to say:

- Real images are real with high probability.
- Fake images are real with low probability.



(Real image)

Discriminator

95% probability of being real

Discriminator wants this to be high

Generator

Discriminator

32% probability of being real

Discriminator wants this to be low

GANs: Training the Discriminator

Which loss does this remind you of?

Discriminator wants to maximize:

$$\underbrace{E_x\left[log(D(x))\right]} + \underbrace{E_z\left[log(1 - D(G(z)))\right]}$$

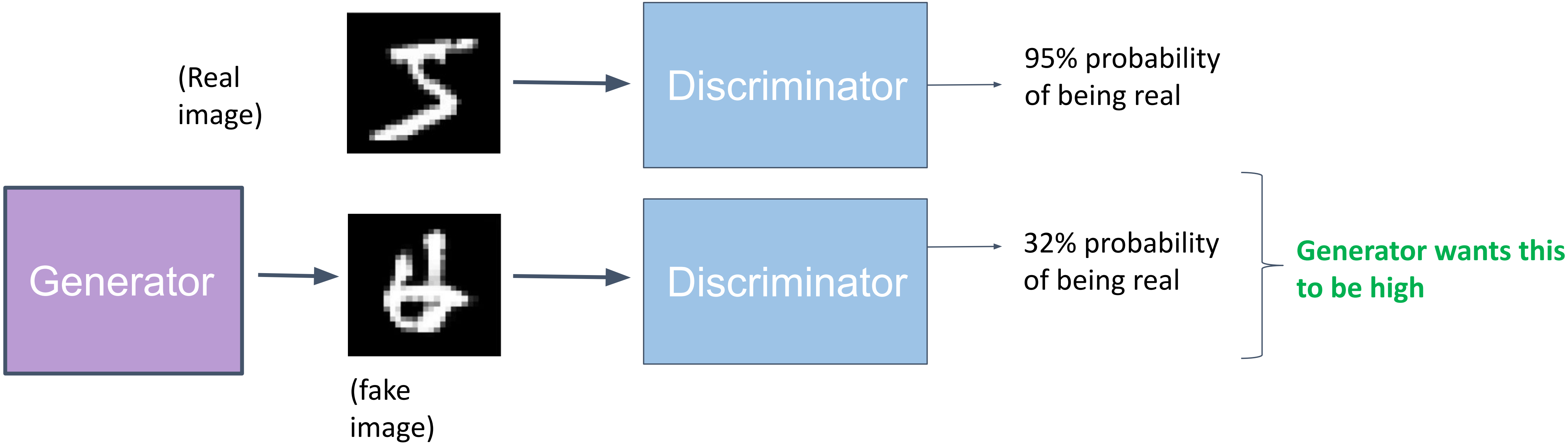
Log probability that the real image x is predicted to be real by the discriminator.

Log probability that the fake image $G(z)$ is predicted to be fake by the discriminator.

Note: Maximizing this quantity is equivalent to minimizing binary cross entropy loss with fake data labelled as 0 and real data labelled as 1.

GANs: Training the Generator

Generator wants to fool the discriminator. It wants the probability of the discriminator saying a fake image is real to be high.



(Real image)

Discriminator

95% probability of being real

Generator

(fake image)

Discriminator

32% probability of being real

Generator wants this to be high

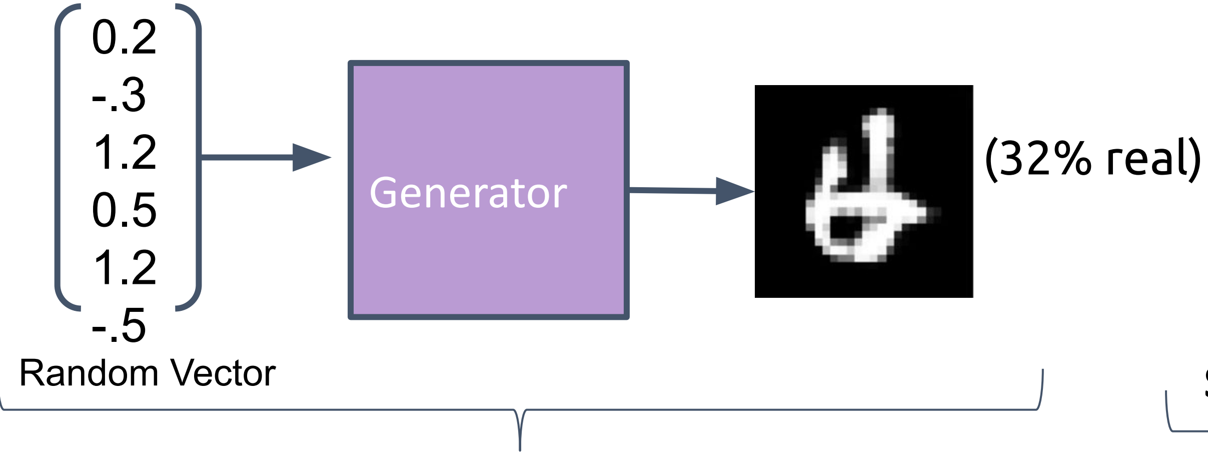
GANs: Training the Generator

Generator wants to maximize:

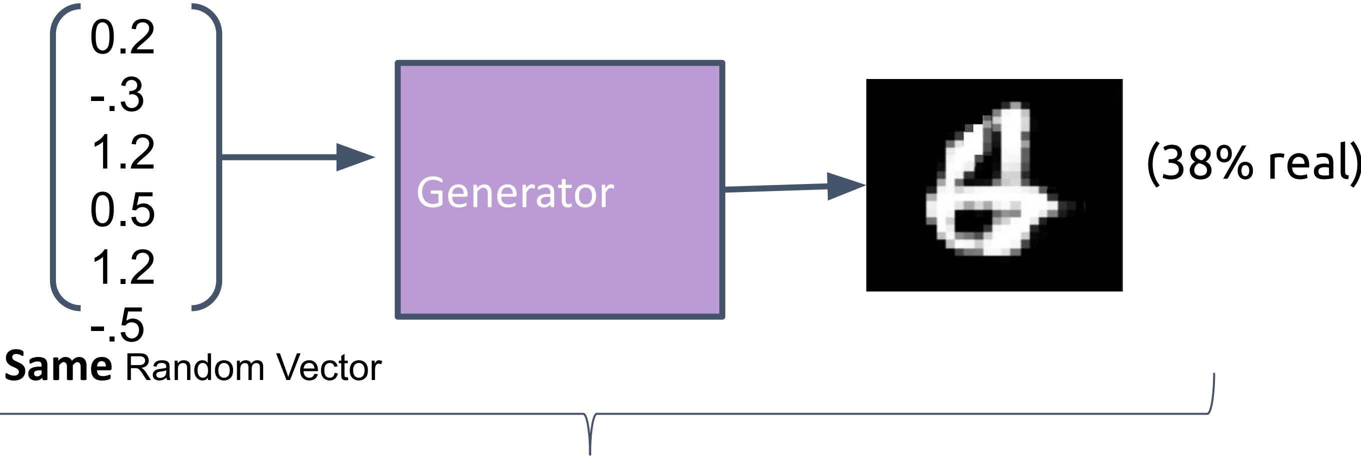
$$E_z[log(D(G(z)))]$$

Log probability that the fake image z is predicted as real by the discriminator.

The generator is only allowed to change ***its own weights*** to maximize this value. Performing an update on the generator will cause all of the images to become slightly more realistic according to the discriminator.



Before weight update



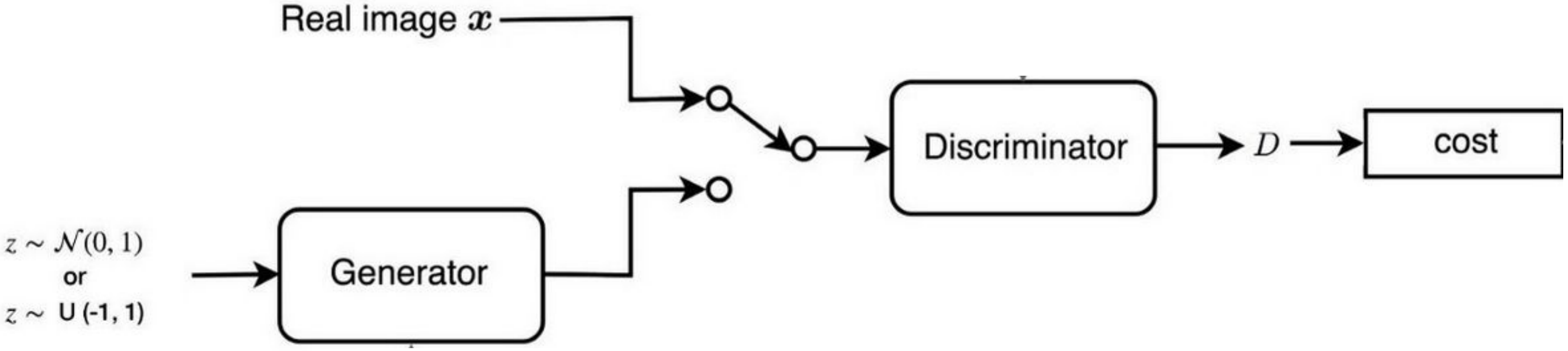
After weight update

GAN Loss

$$E_x[log(D(x))] + E_z[log(1 - D(G(z)))]$$

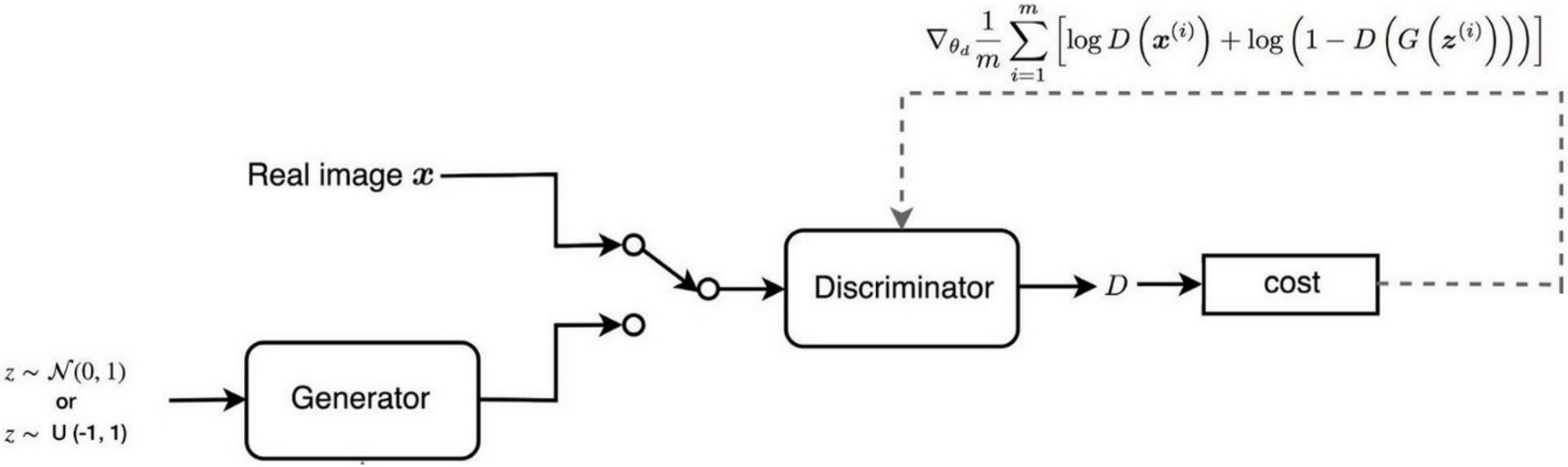
https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

GAN Loss



https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

GAN Loss



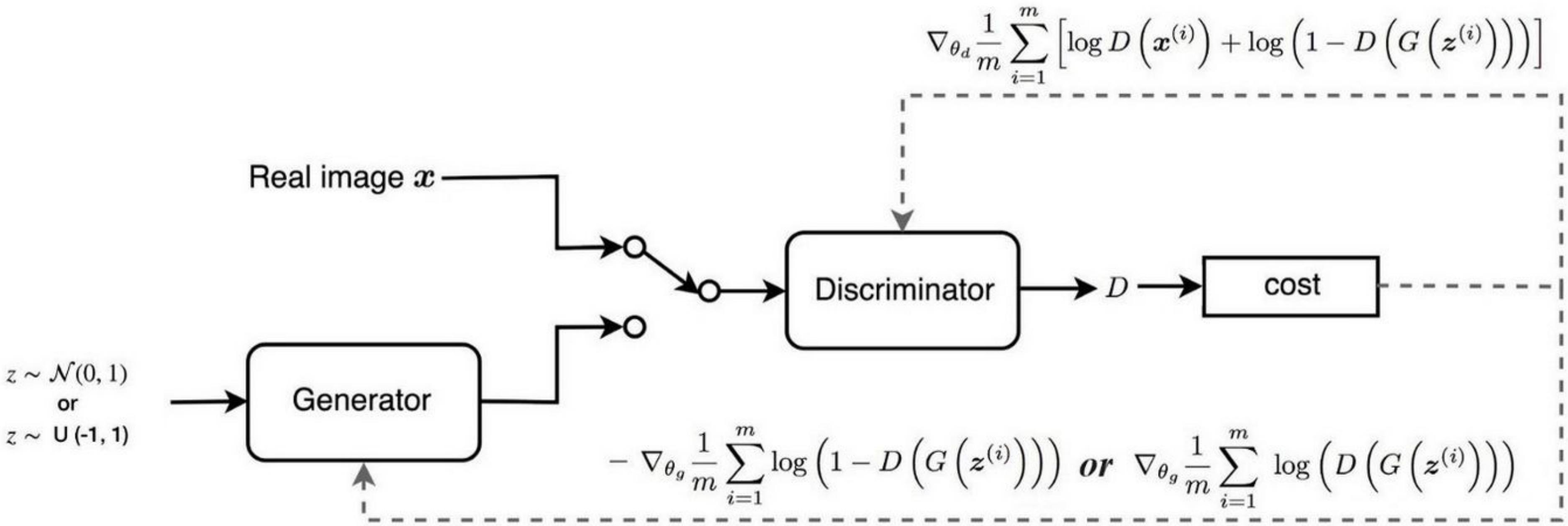
$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$
Real image $\boldsymbol{x}$
Discriminator
D
cost
$z \sim \mathcal{N}(0,1)$
or
$z \sim$ U (-1, 1)
Generator

What about generator?

https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

GAN Loss

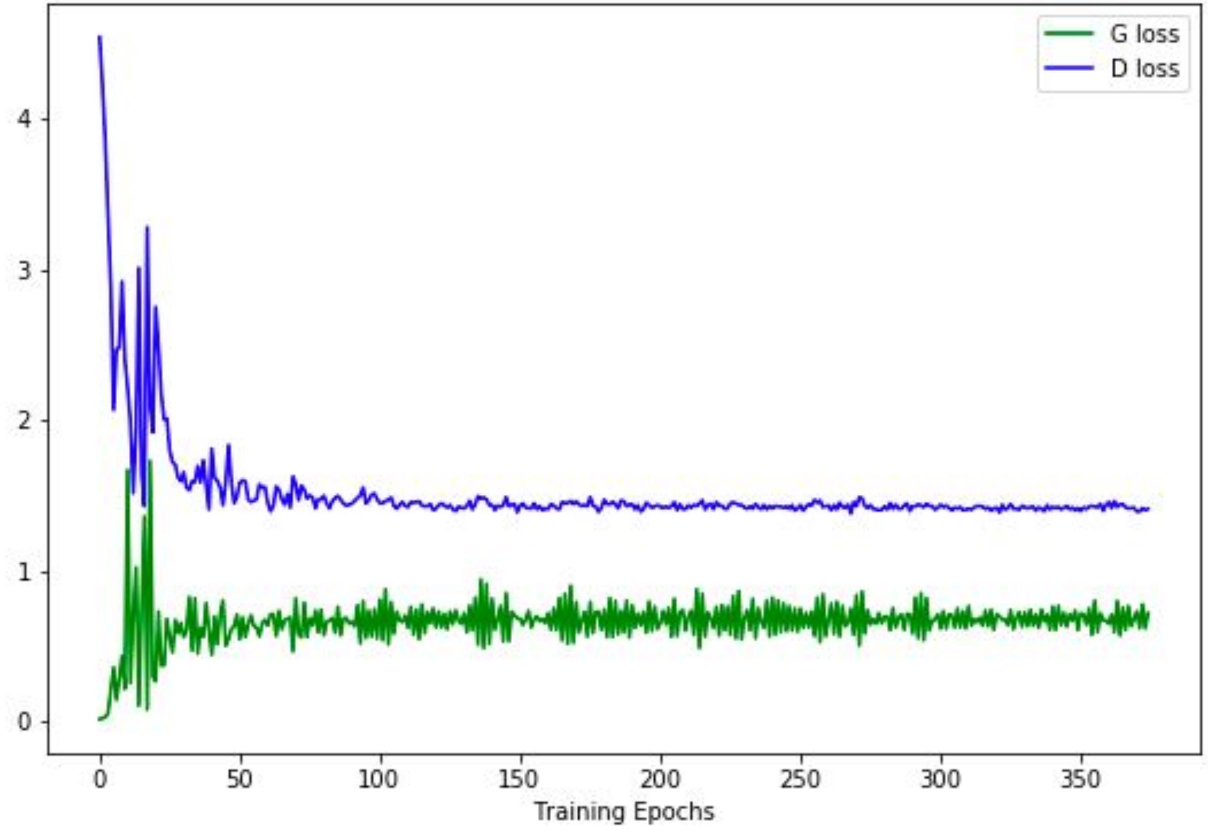
Any questions?



$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$
Real image $\boldsymbol{x}$
Discriminator
D
cost
$z \sim \mathcal{N}(0, 1)$
or
$z \sim$ U (-1, 1)
Generator
$-\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$ or $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$

https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

GAN Training Dynamics



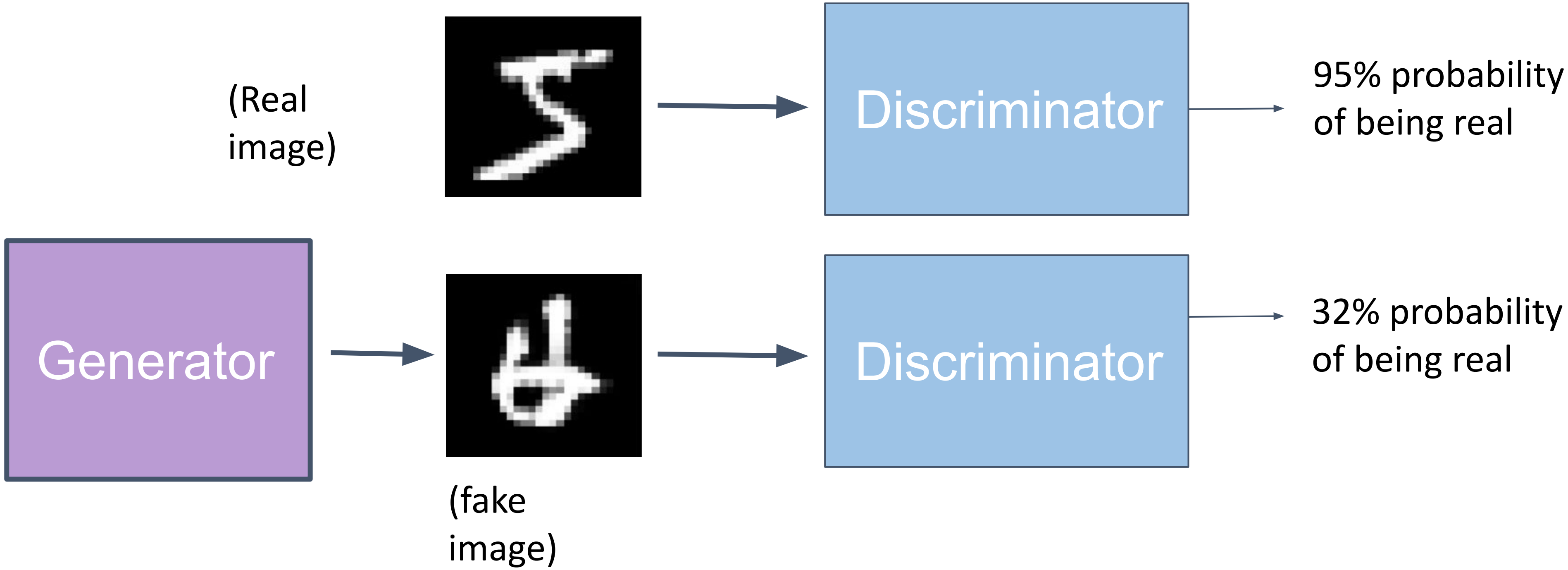
- Does not exhibit the typical "training loss continues to go down" behavior
- ***Why?***
  - Training a GAN is a "stalemate" – G and D continually adjust to each other's improvements
  - More formally, training a GAN to convergence is attempting to find an equilibrium of a two-player minimax game

Demo

https://poloclub.github.io/ganlab/

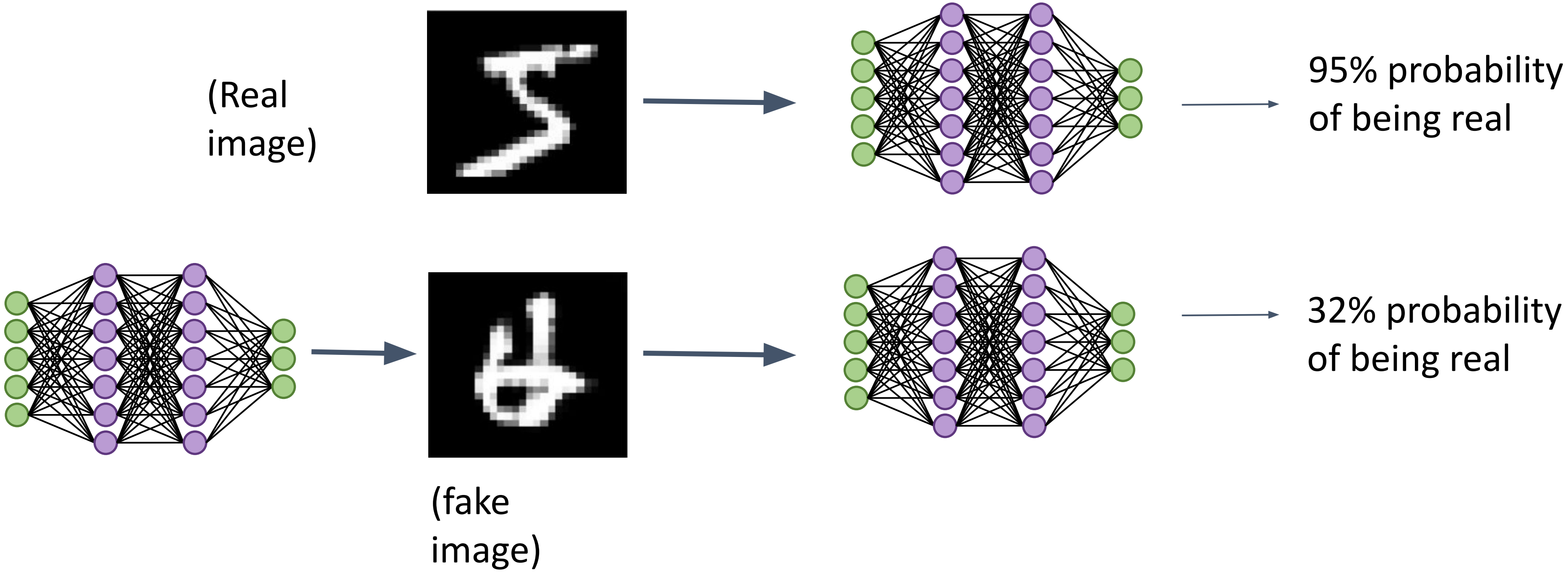
What do G and D look like inside?

- Architecture of the networks determined by problem



What do G and D look like inside?

- Architecture of the networks determined by problem
- **Fully connected**



(Real image)

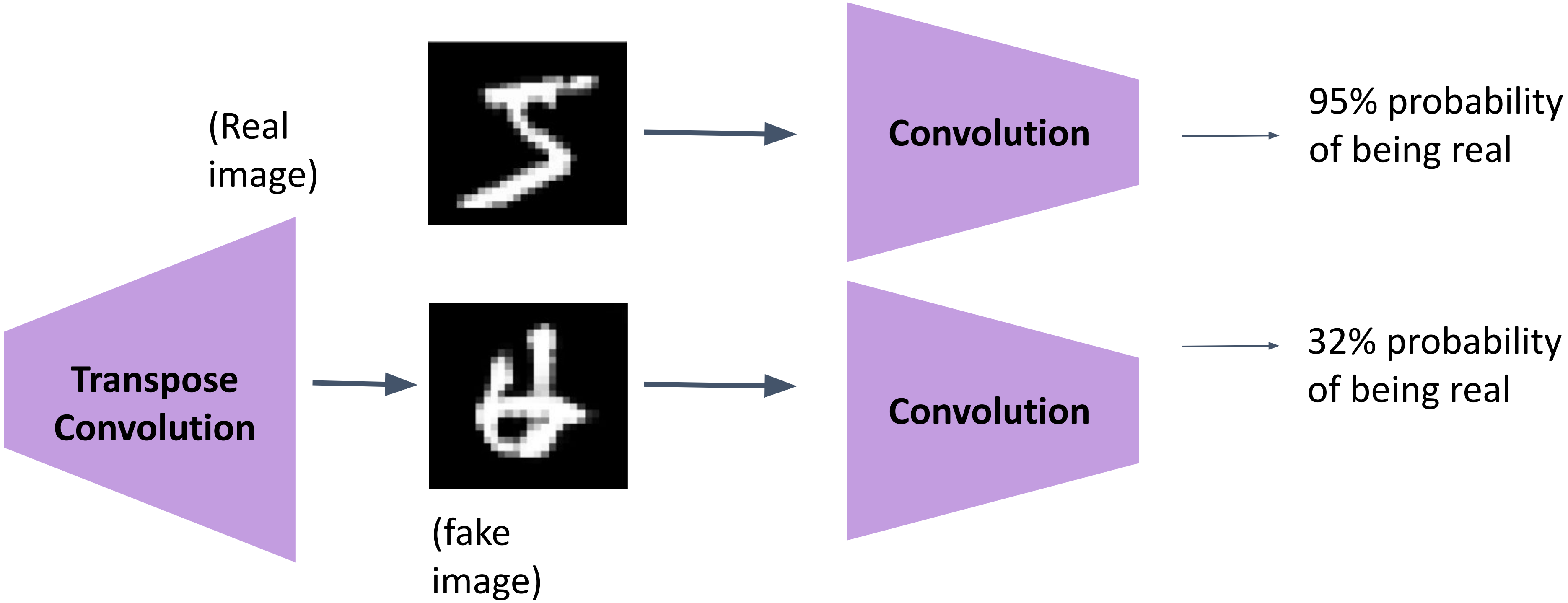
95% probability of being real

32% probability of being real

(fake image)

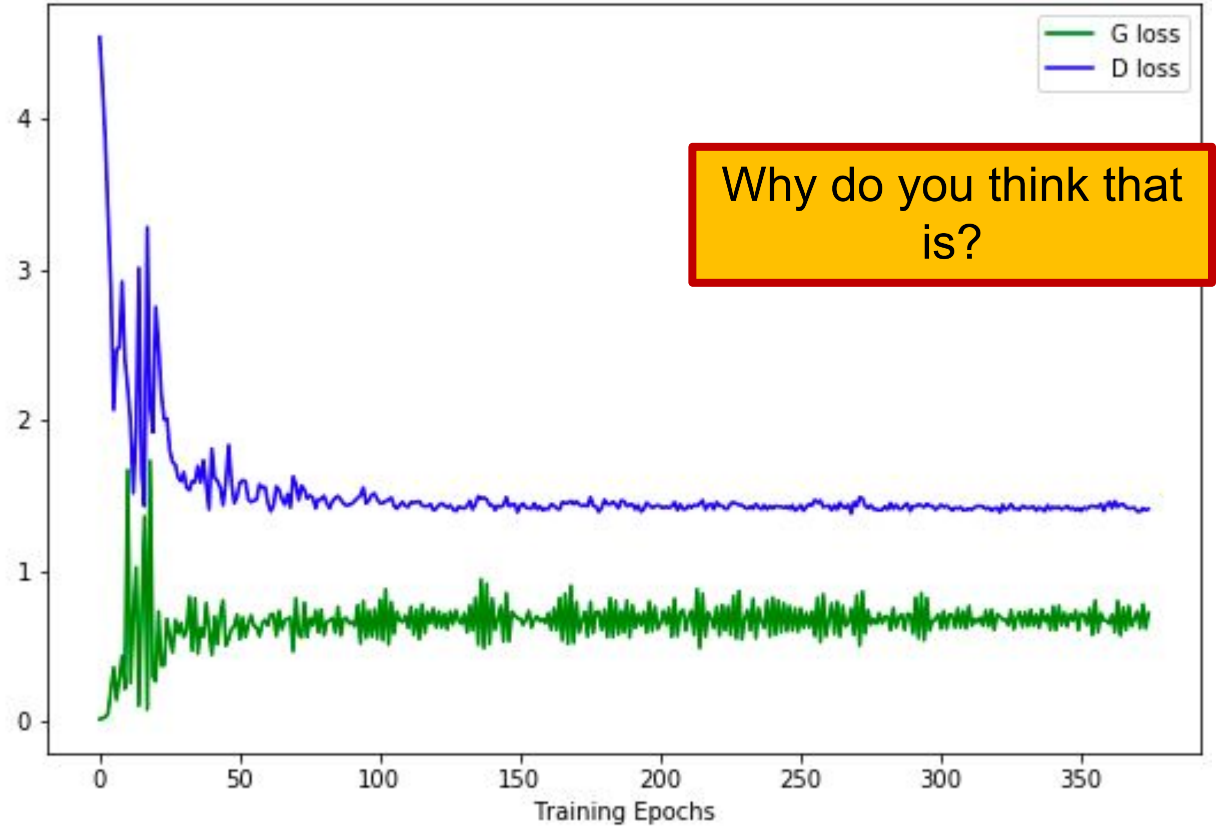
What do G and D look like inside?

- Architecture of the networks determined by problem
- **Convolutional / Transpose convolutional**



Problems with GANs

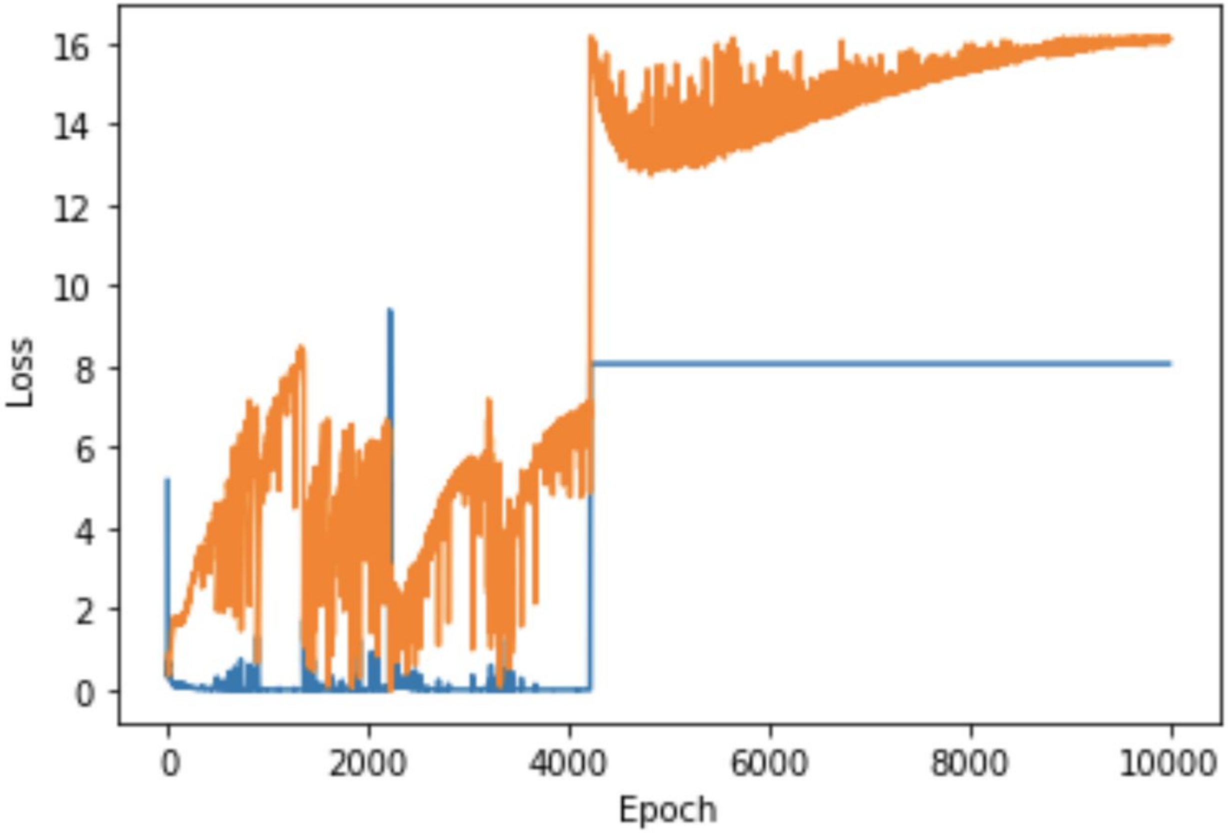
GAN training can be ***very*** unstable



Why do you think that is?

- This picture? You get this if everything is working well
- Turns out, equilibria are hard to find
  - With every other net we've trained, the loss function is with respect to a fixed target value we're trying to hit
  - Here, we have a "moving target" (G's target is fool D, D's target is detect G)
- These curves can oscillate a lot

GAN training can be ***very*** unstable



- In particular: what happens if the discriminator ever becomes perfect at detecting G's fakes?
  - The discriminator always returns probability zero
  - Since D is returning a constant, the gradient through D is zero
  - The generator stops training

Vanishing gradient

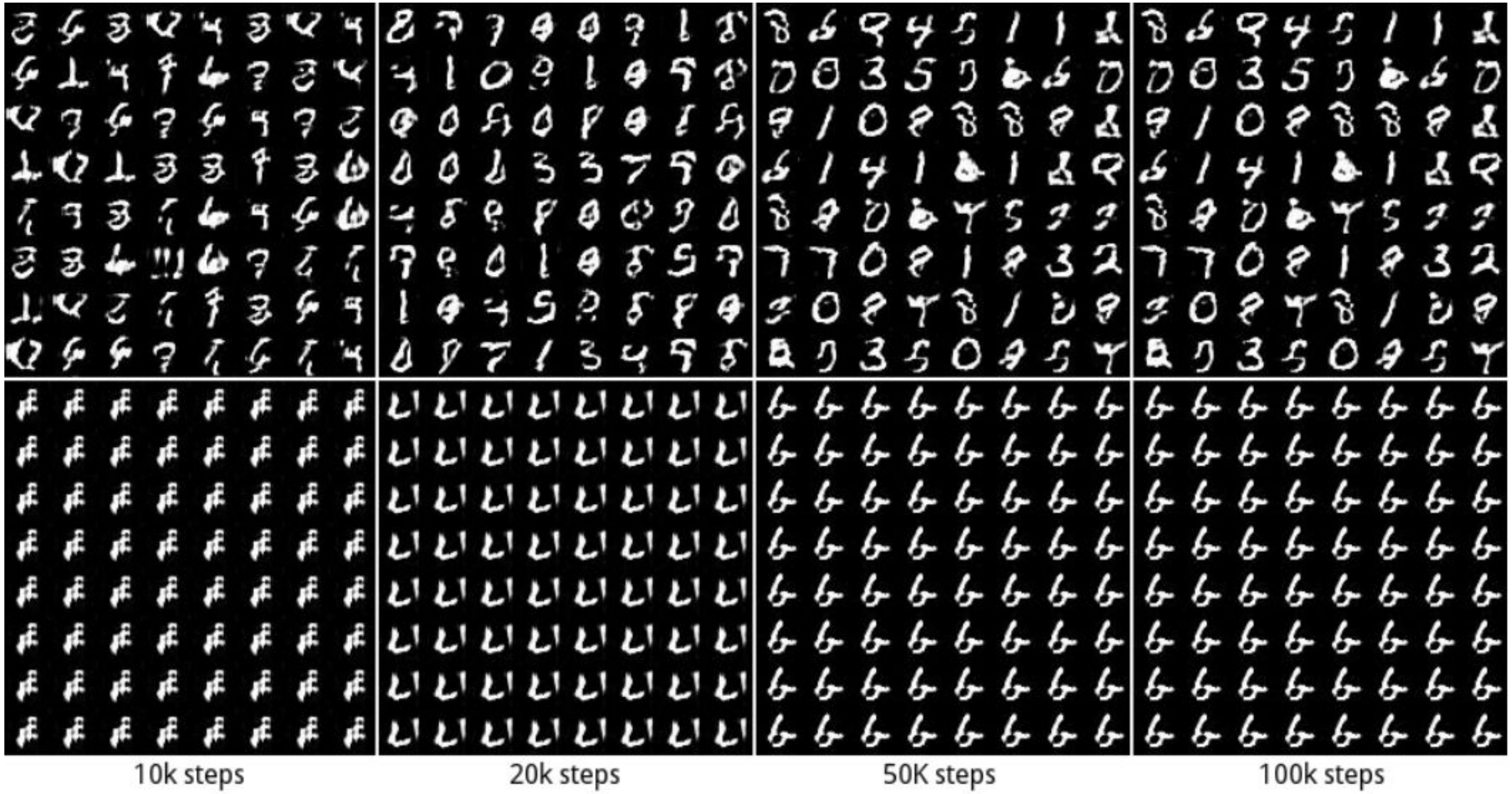
Generator loss: $E_z[log(D(G(z)))]$

Mode Collapse

- Generator loss says: “generate an output that looks real”
- It does not say: “generate ***every*** output that looks real”
- The generator can “cheat” by finding one output / a few outputs that reliably fool the discriminator (the specific one(s) it finds can shift over training)

How do we fix this?

Mode Collapse

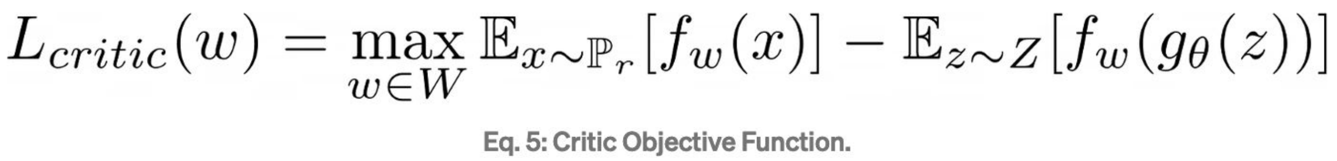


Output from a healthy GAN

Output from a GAN with mode collapse. All outputs from GAN, regardless of random input noise, are the same.

https://arxiv.org/pdf/1611.02163.pdf

Wasserstein GANs (WGANs)

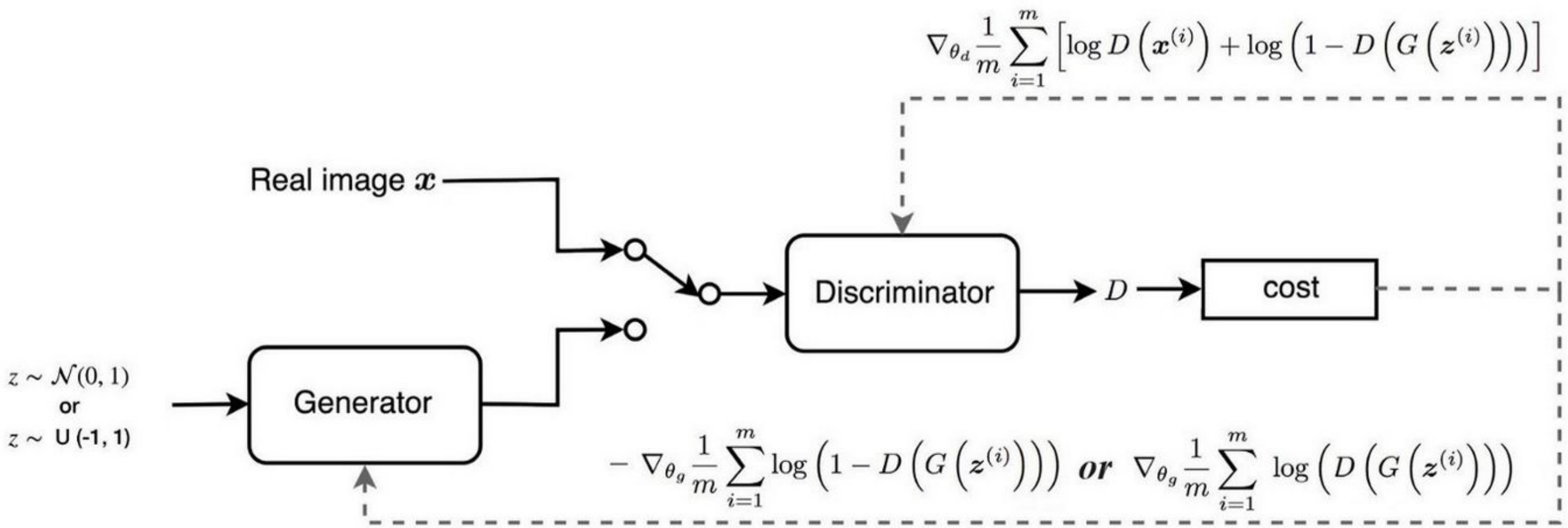


$$L_{critic}(w) = \max_{w \in W} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim Z}[f_w(g_\theta(z))]$$

Eq. 5: Critic Objective Function.

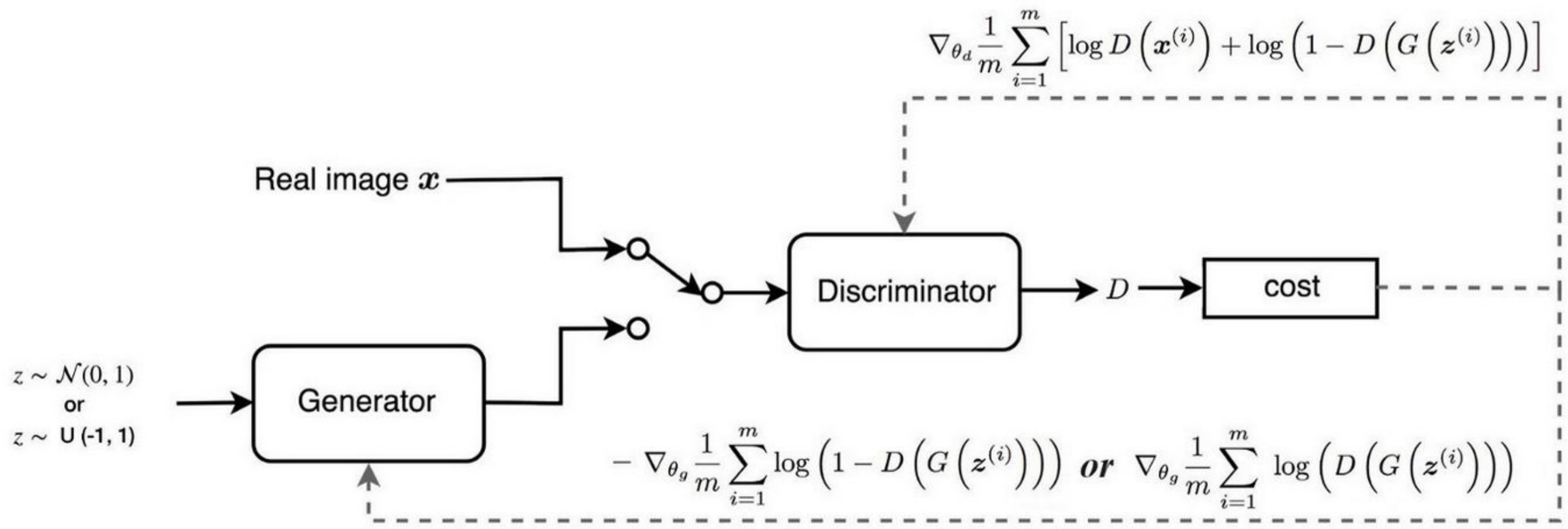
https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

GAN

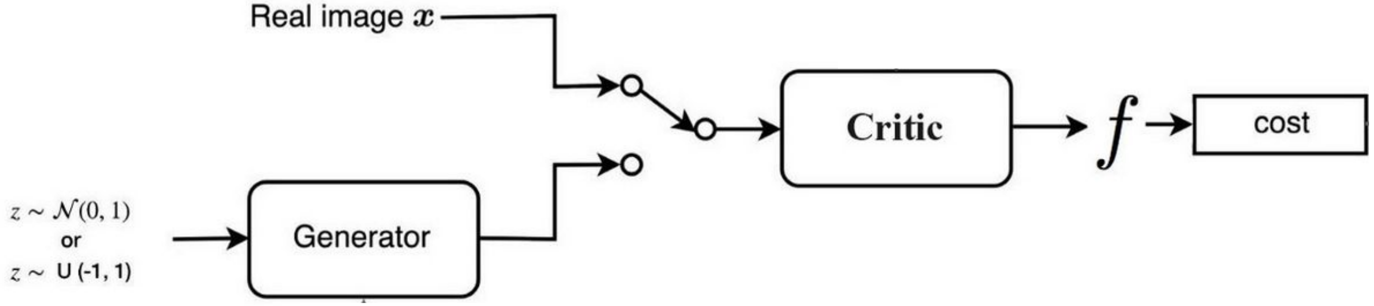

$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$
Real image $\boldsymbol{x}$
Discriminator
D
cost
$z \sim \mathcal{N}(0,1)$
or
$z \sim$ U (-1, 1)
Generator
$-\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$ ***or*** $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$

https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

GAN

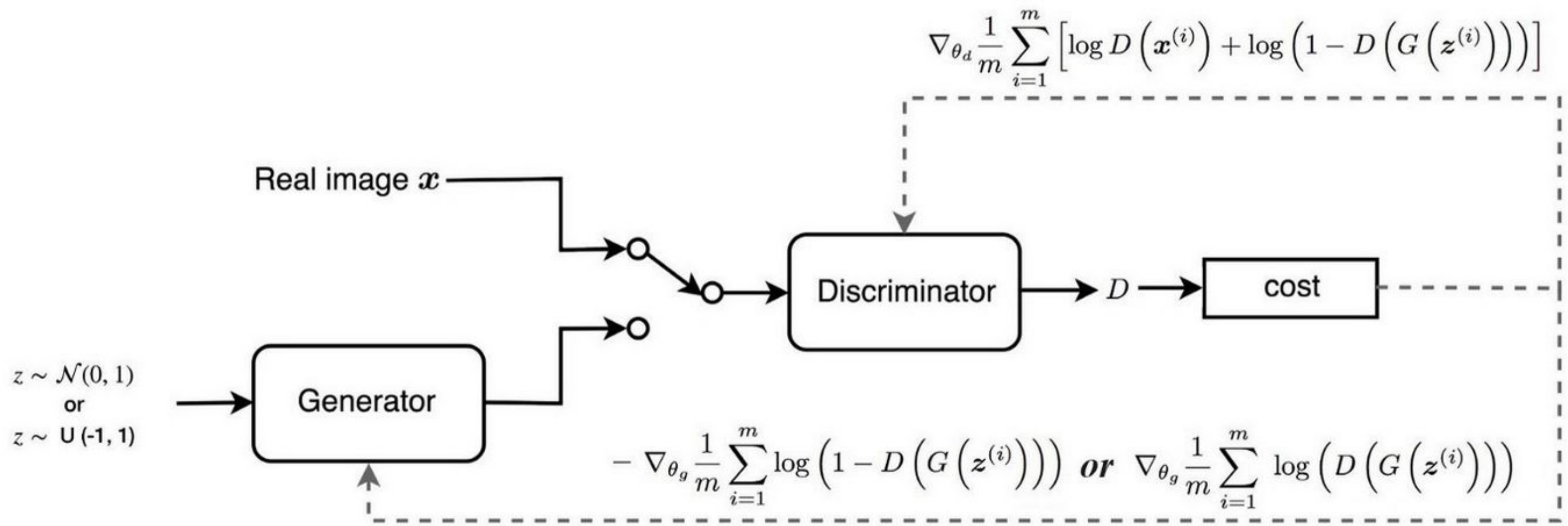

$\nabla_{\theta_d} \frac{1}{m}\sum_{i=1}^{m}\left[\log D\left(\boldsymbol{x}^{(i)}\right)+\log\left(1-D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$
Real image $\boldsymbol{x}$
Discriminator
D
cost
$z \sim \mathcal{N}(0,1)$
or
$z \sim$ U (-1, 1)
Generator
$-\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log\left(1-D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$ or $\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log\left(D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$

WGAN

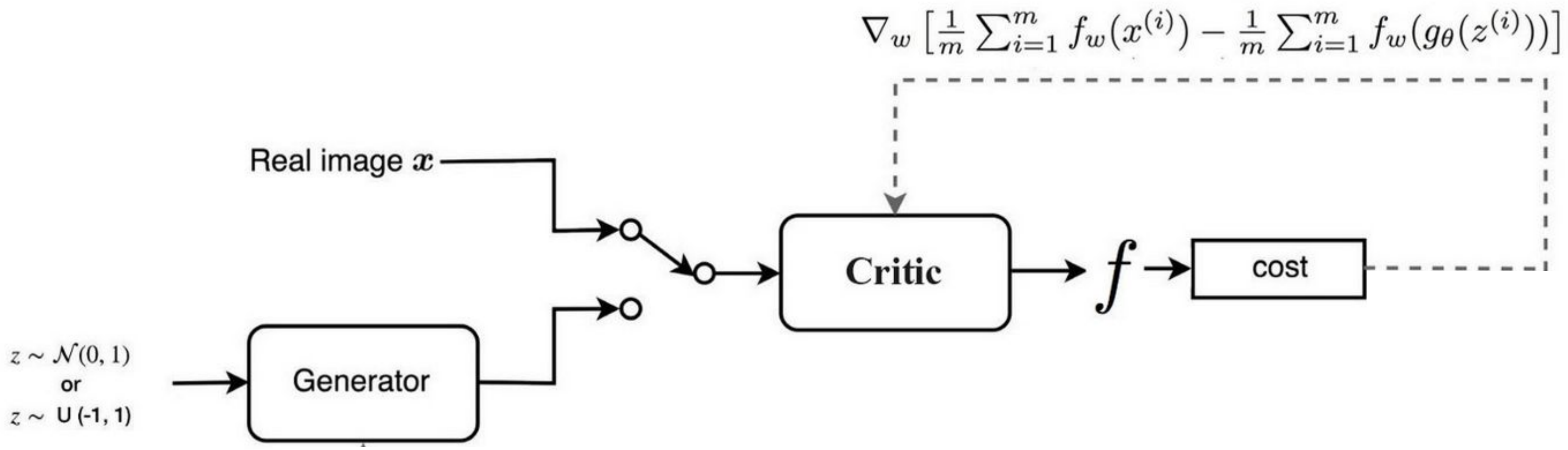


https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

GAN


$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$
Real image $\boldsymbol{x}$
Discriminator
D
cost
$z \sim \mathcal{N}(0, 1)$
or
$z \sim$ U (-1, 1)
Generator
$-\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$ ***or*** $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$

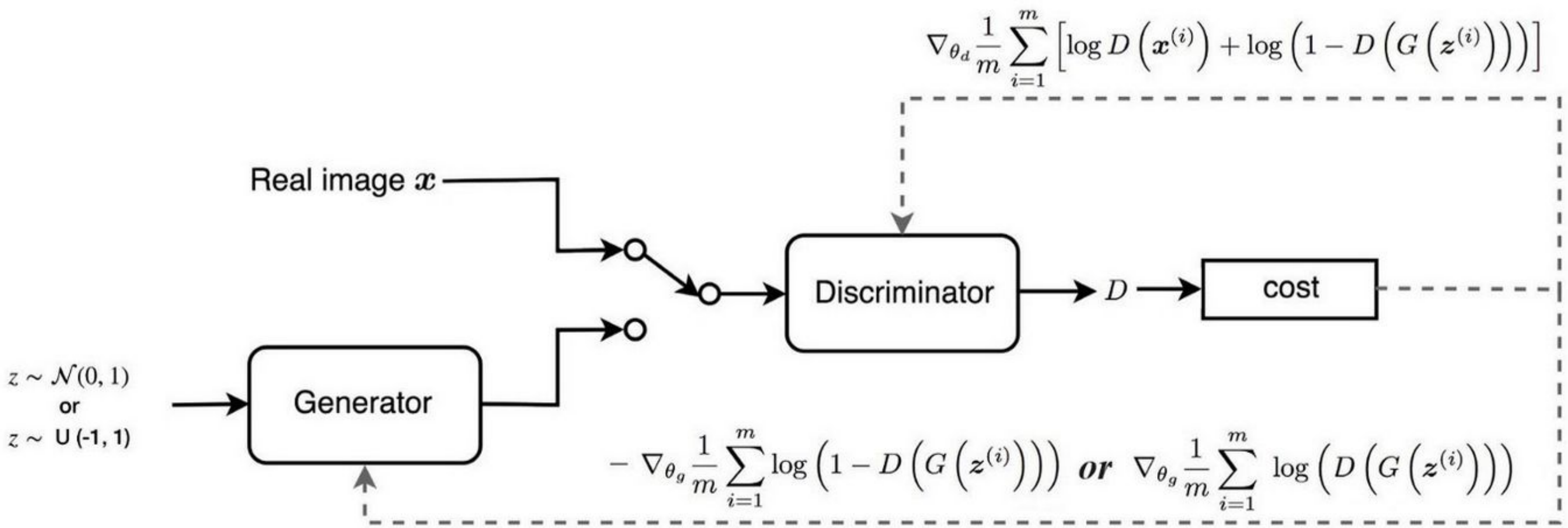
WGAN


$\nabla_w \left[\frac{1}{m} \sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m} \sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
Real image $\boldsymbol{x}$
Critic
f
cost
$z \sim \mathcal{N}(0, 1)$
or
$z \sim$ U (-1, 1)
Generator

https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

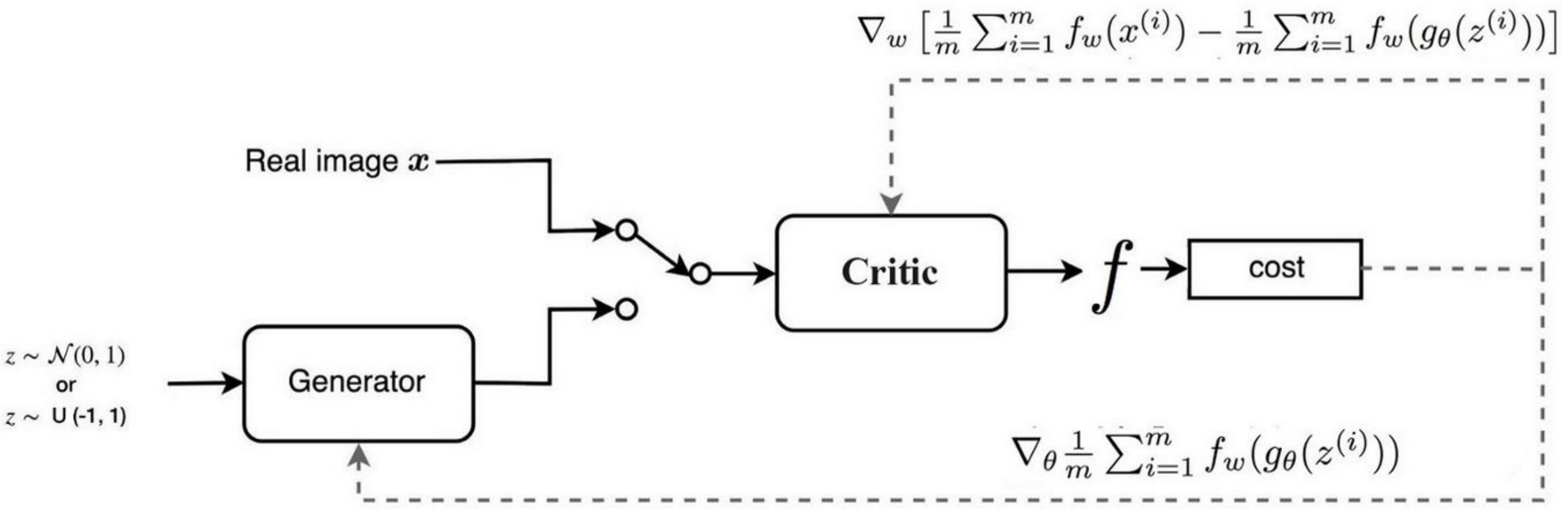
Any questions?

GAN



$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$
Real image $\boldsymbol{x}$
Discriminator
D
cost
$z \sim \mathcal{N}(0, 1)$
or
$z \sim$ U (-1, 1)
Generator
$-\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$ *or* $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$

WGAN



$\nabla_w \left[\frac{1}{m} \sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m} \sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
Real image $\boldsymbol{x}$
Critic
f
cost
$z \sim \mathcal{N}(0, 1)$
or
$z \sim$ U (-1, 1)
Generator
$\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$

https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

Diffusion models

- State-of-the-art models for image generation

Stable.AI DALL·E 2

- Guest lectures by Calvin Luo (CS Ph.D. student) – Wednesday and Friday this week

Today’s goal – learn about generative adversarial networks (GANs)

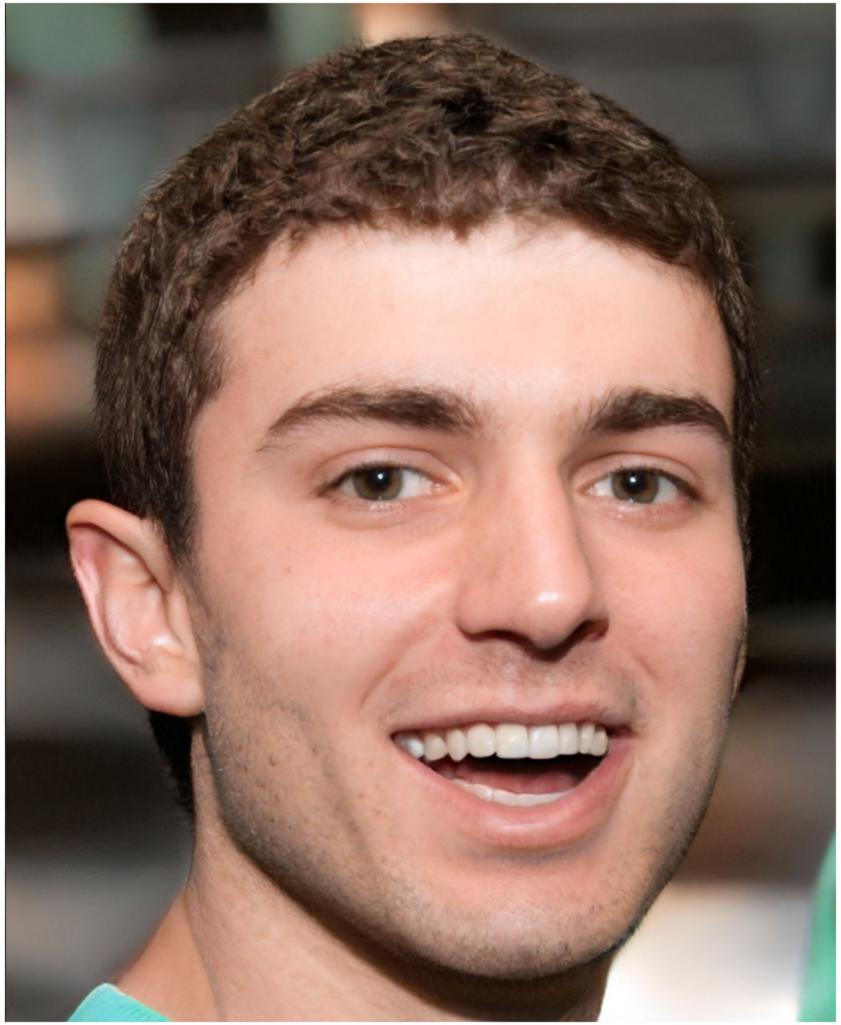
(1) Generative Adversarial Networks (GANs)

(2) Training GANs and challenges

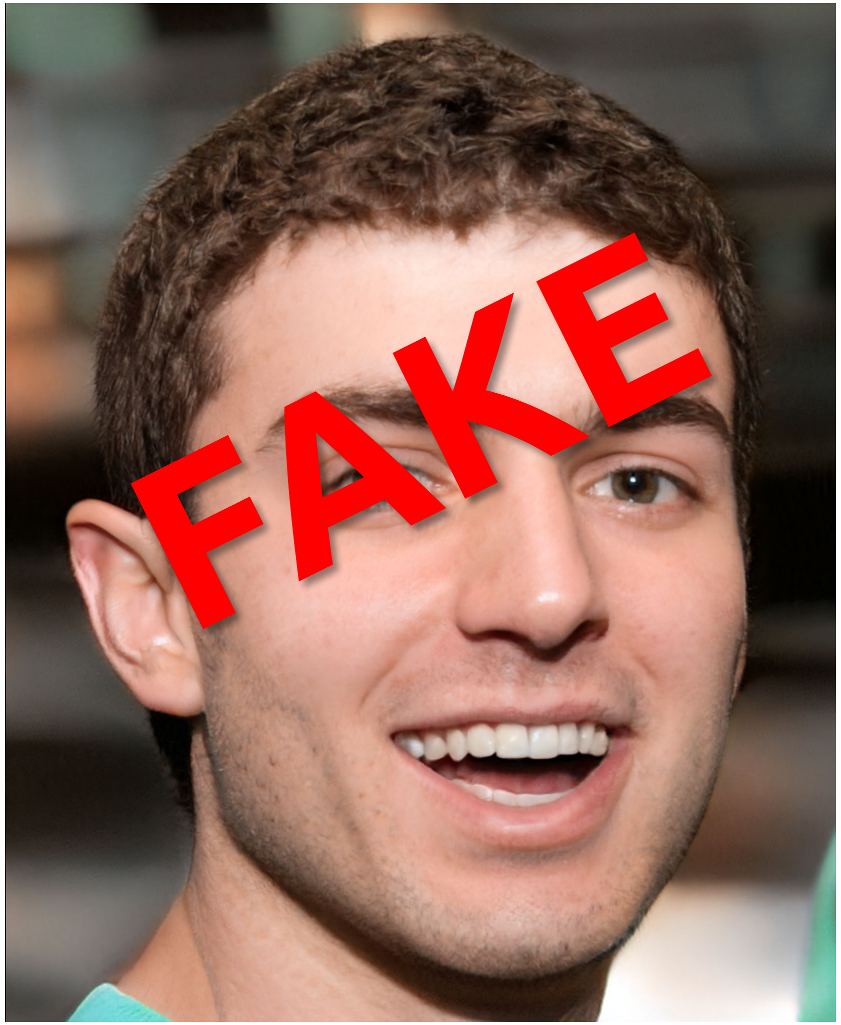
(3) Deepfakes

Deep generative models are getting really good

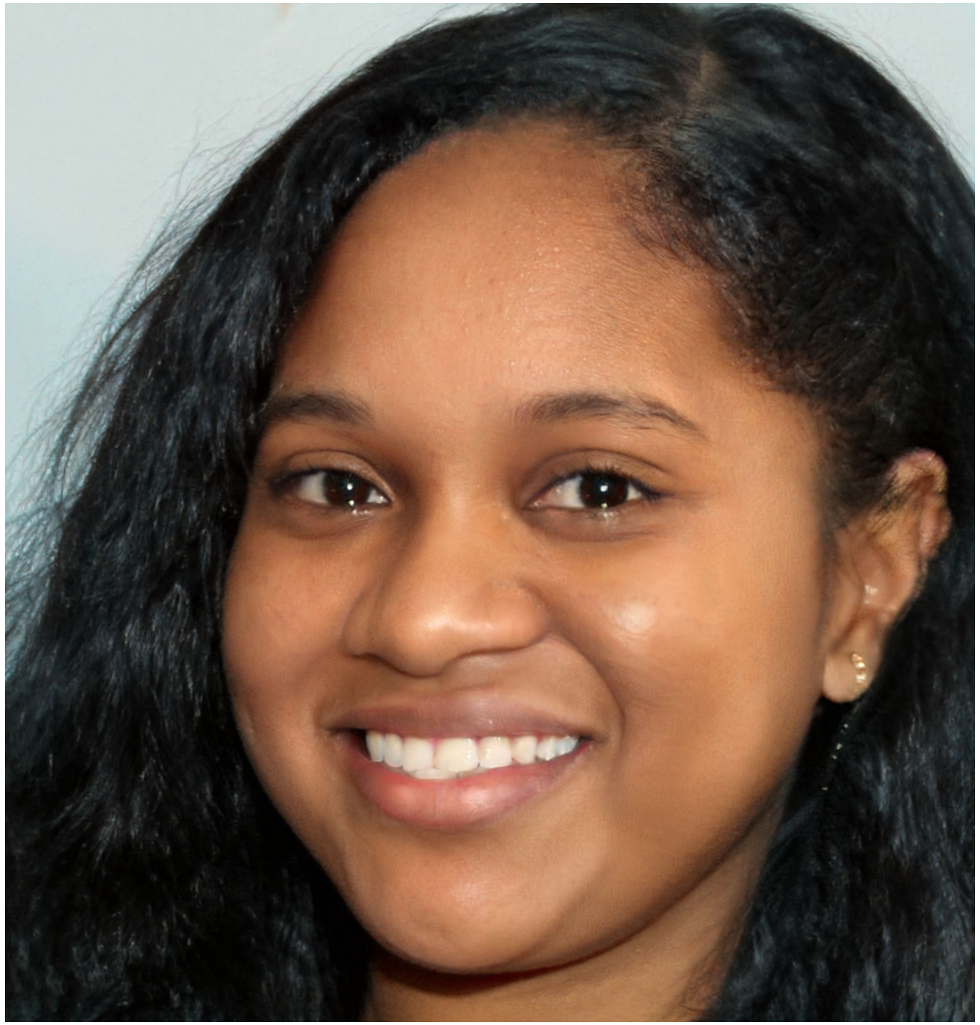
Is this image real or generated?



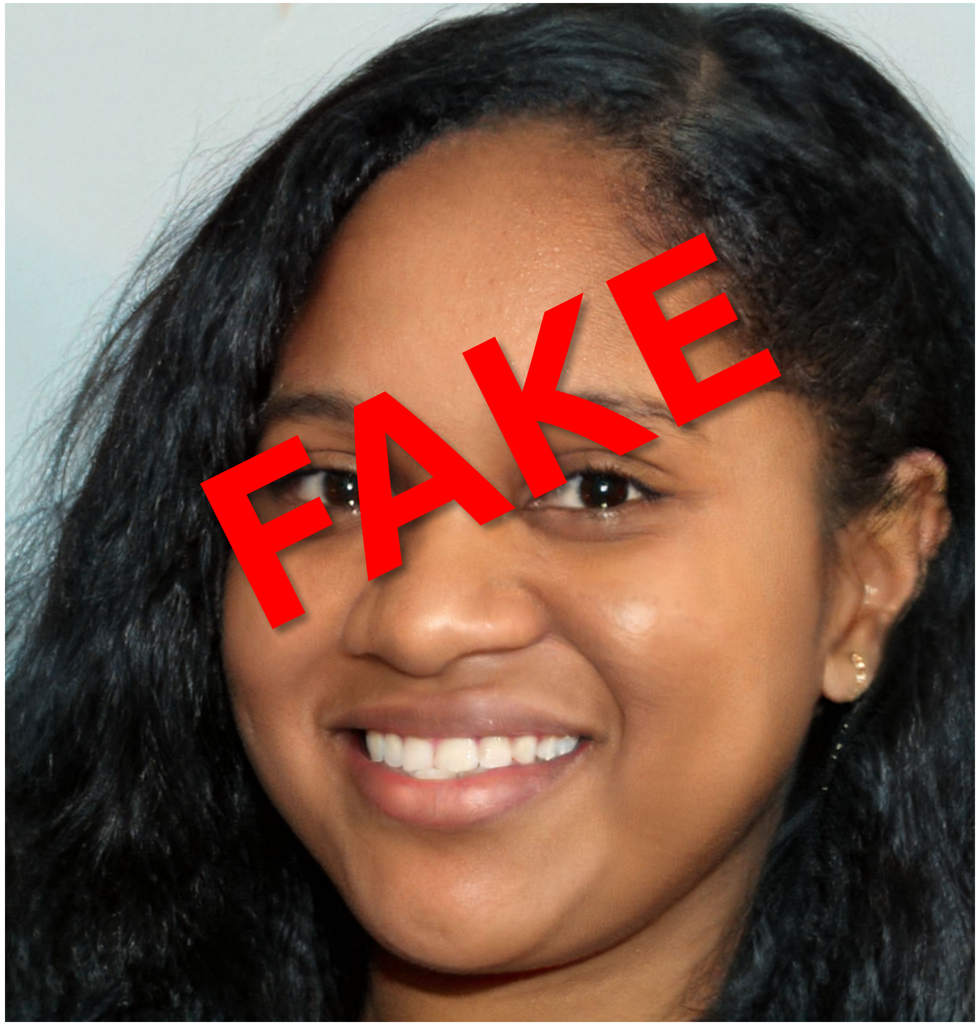
Is this image real or generated?



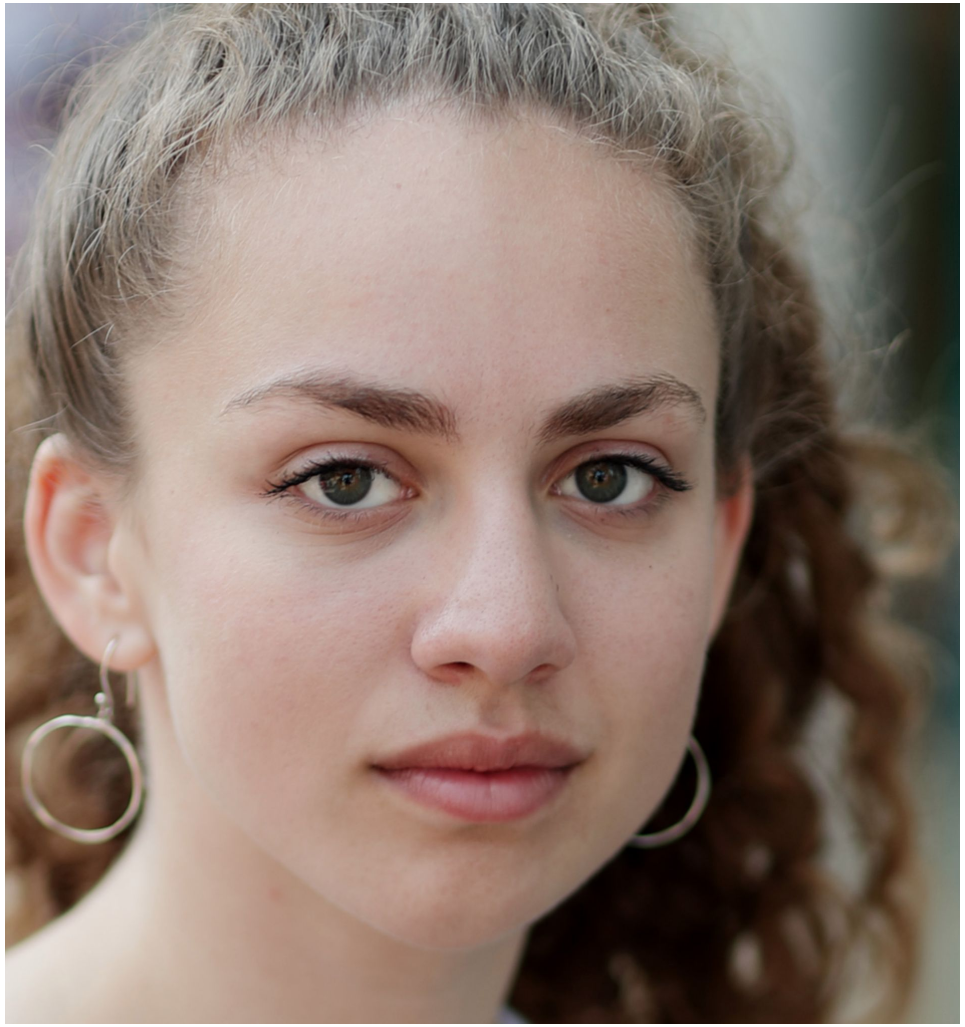
Is this image real or generated?



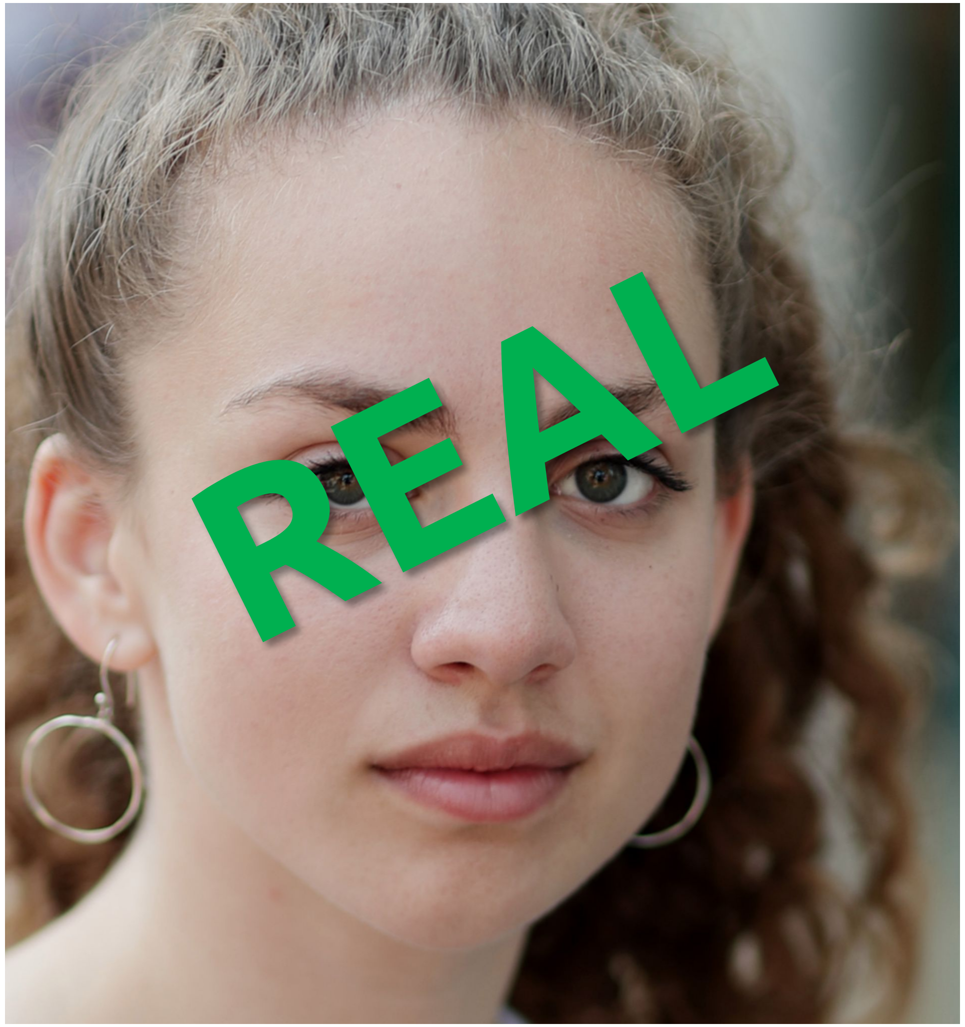
Is this image real or generated?



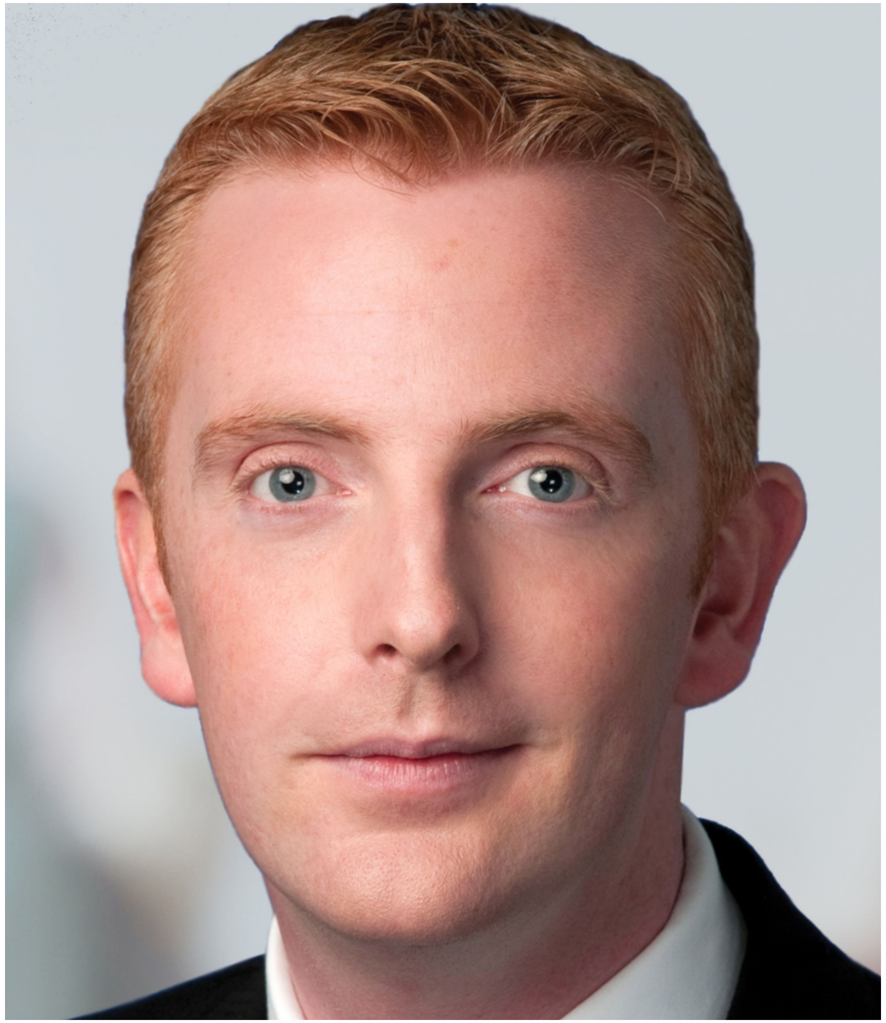
Is this image real or generated?



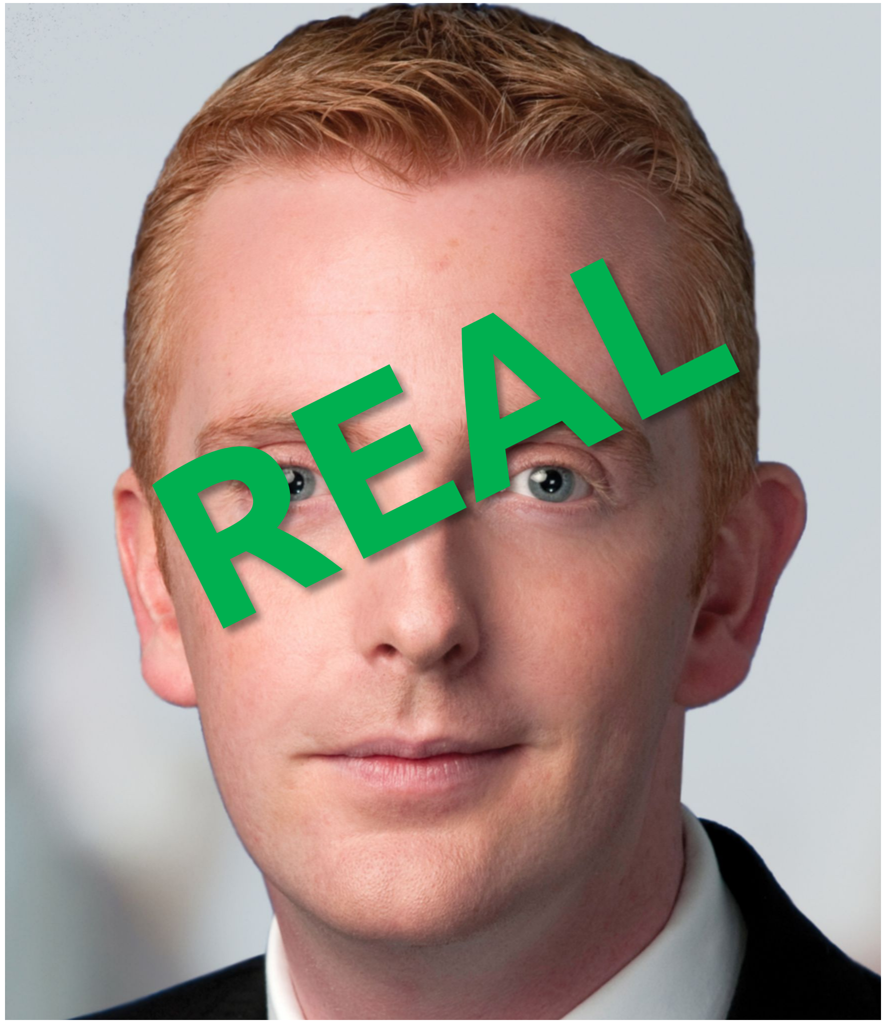
Is this image real or generated?



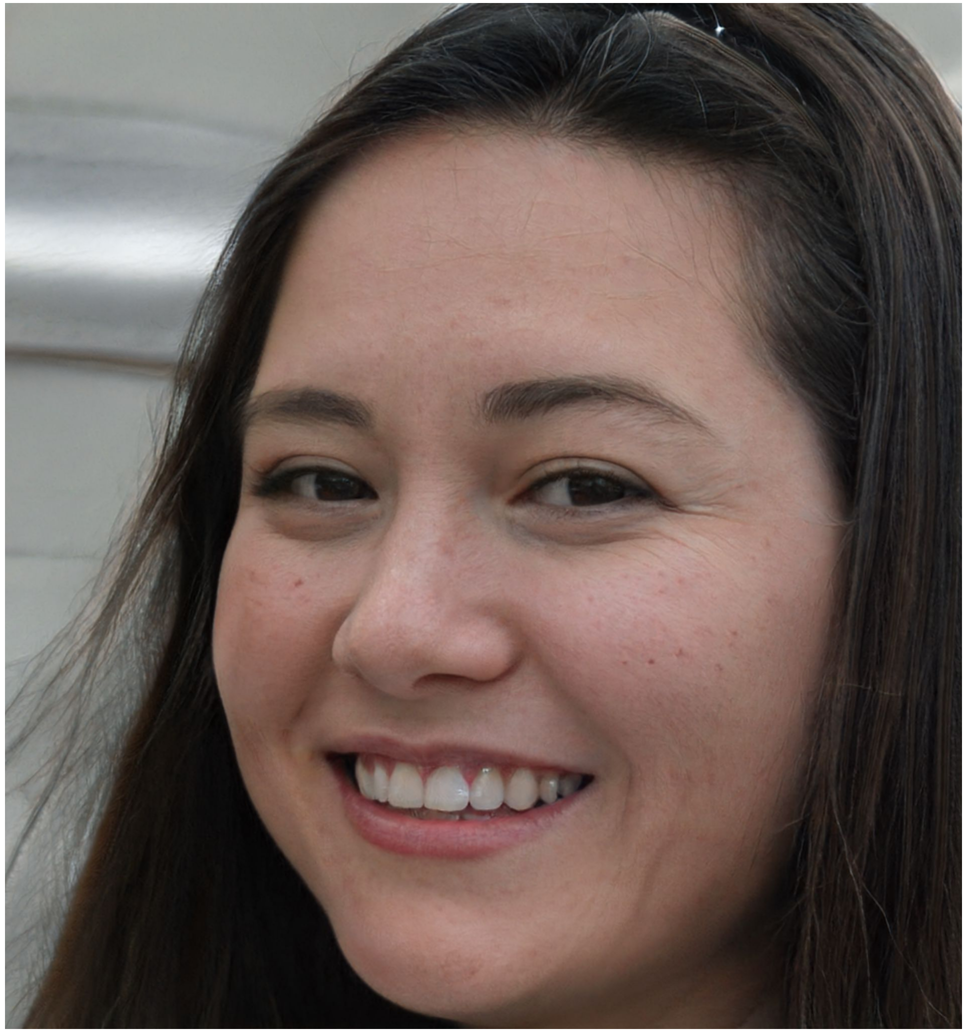
Is this image real or generated?



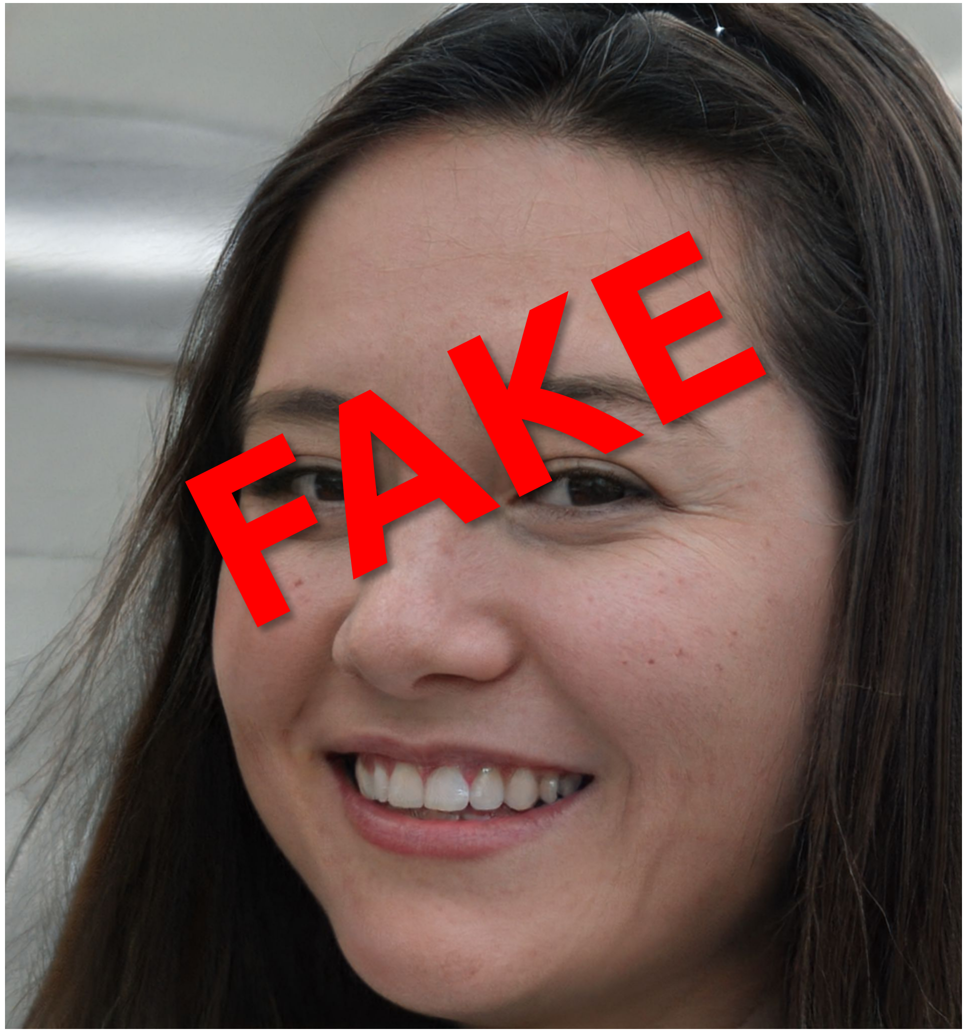
Is this image real or generated?



Is this image real or generated?



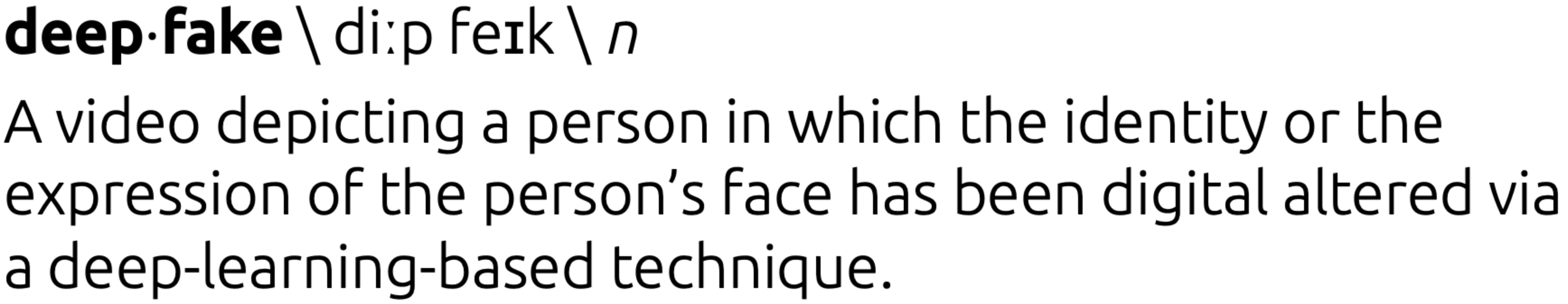
Is this image real or generated?



What is a “deep fake?”

For the purposes of this class:

-

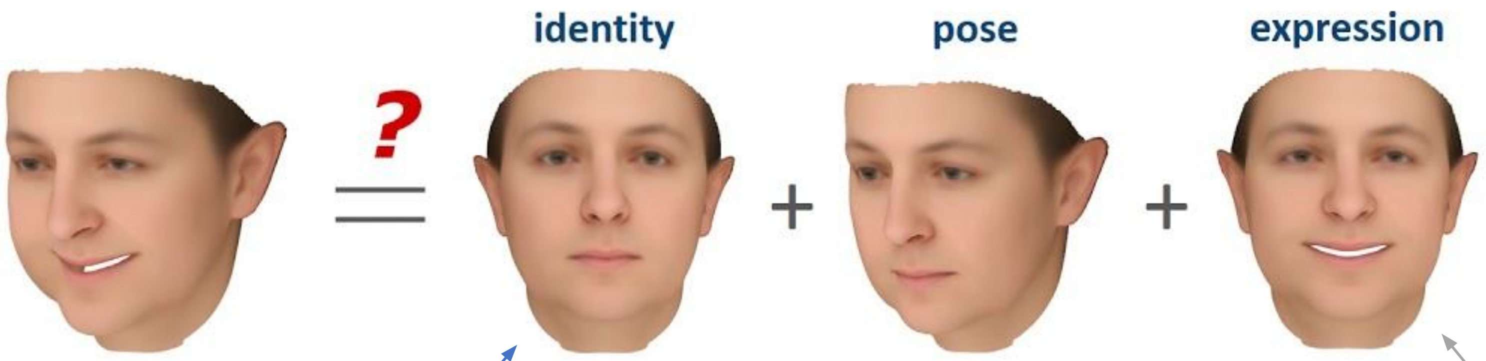


deep·fake \ diːp feɪk \ *n*

A video depicting a person in which the identity or the expression of the person’s face has been digital altered via a deep-learning-based technique.

What kinds of alterations?

- Computer vision researchers use the following scheme to talk about face appearance:



The geometric shape and texture of a person's face

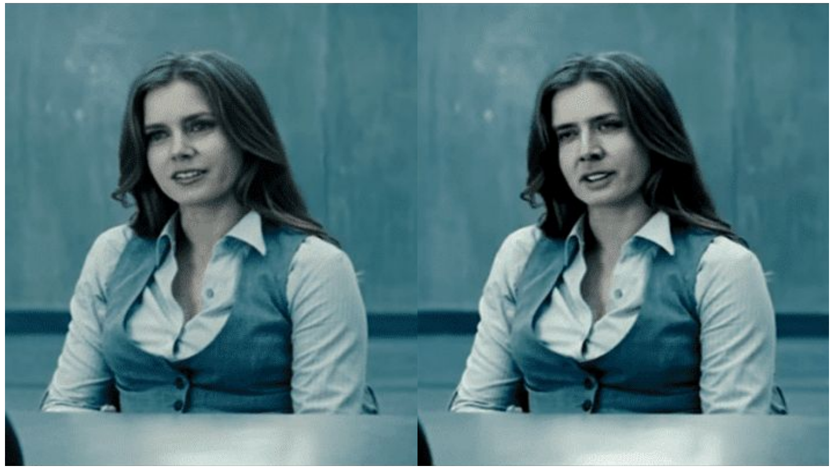
Which way the head is facing (rigid rotation of head; governed by neck muscles)

Facial expression (non-rigid deformation of face; governed by facial muscles)

Can Facial Pose and Expression Be Separated With Weak Perspective Camera?]

Two main “flavors” of deepfake

Face swap



- *Modify identity; keep pose and expression the same*
- **Application**: “digital doubles” (e.g. putting an actor’s face onto a stuntperson’s body)

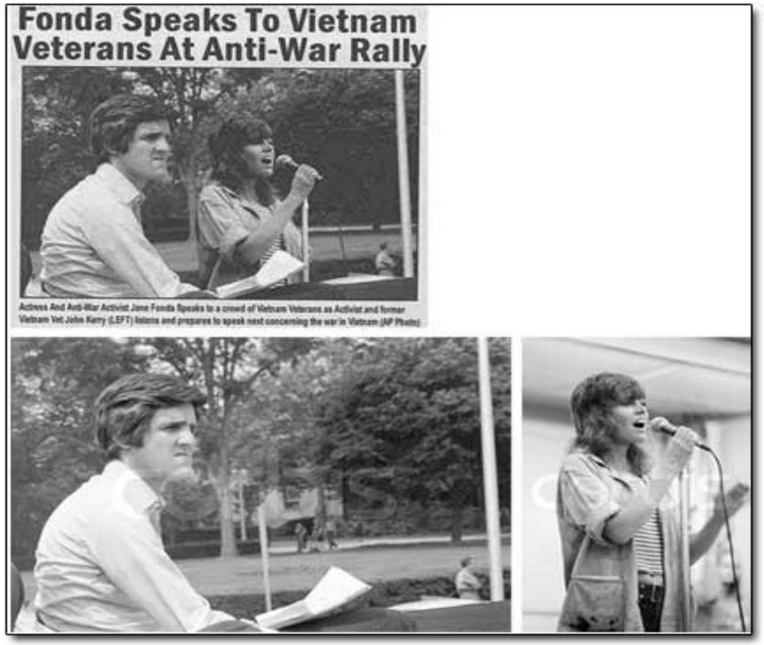
Video puppetry

- *Modify expression (+ pose); keep identity the same*
- **Application:** language dubbing

Why are people worried about deepfakes?

Fake visual media has been around for a while

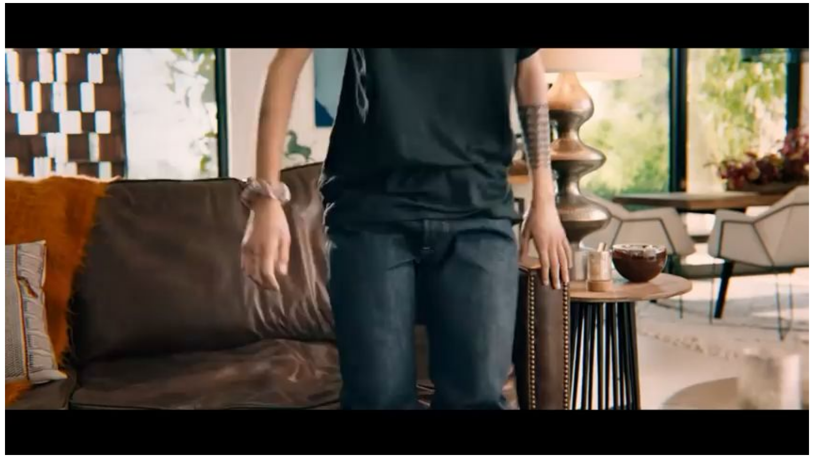
Fake photos



Fonda Speaks To Vietnam Veterans At Anti-War Rally

Actress And Anti-War Activist Jane Fonda Speaks to a crowd of Vietnam Veterans as Activist and former Vietnam Vet John Kerry (LEFT) listens and prepares to speak next concerning the war in Vietnam (AP Photo)

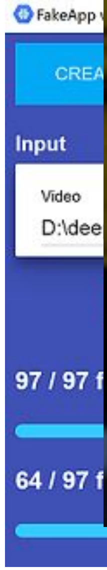
Fake videos



How deepfakes change the game

Ea

- Now any ted (and
  GPU-equ rapidly
  some fre kes
  relatively



FakeApp

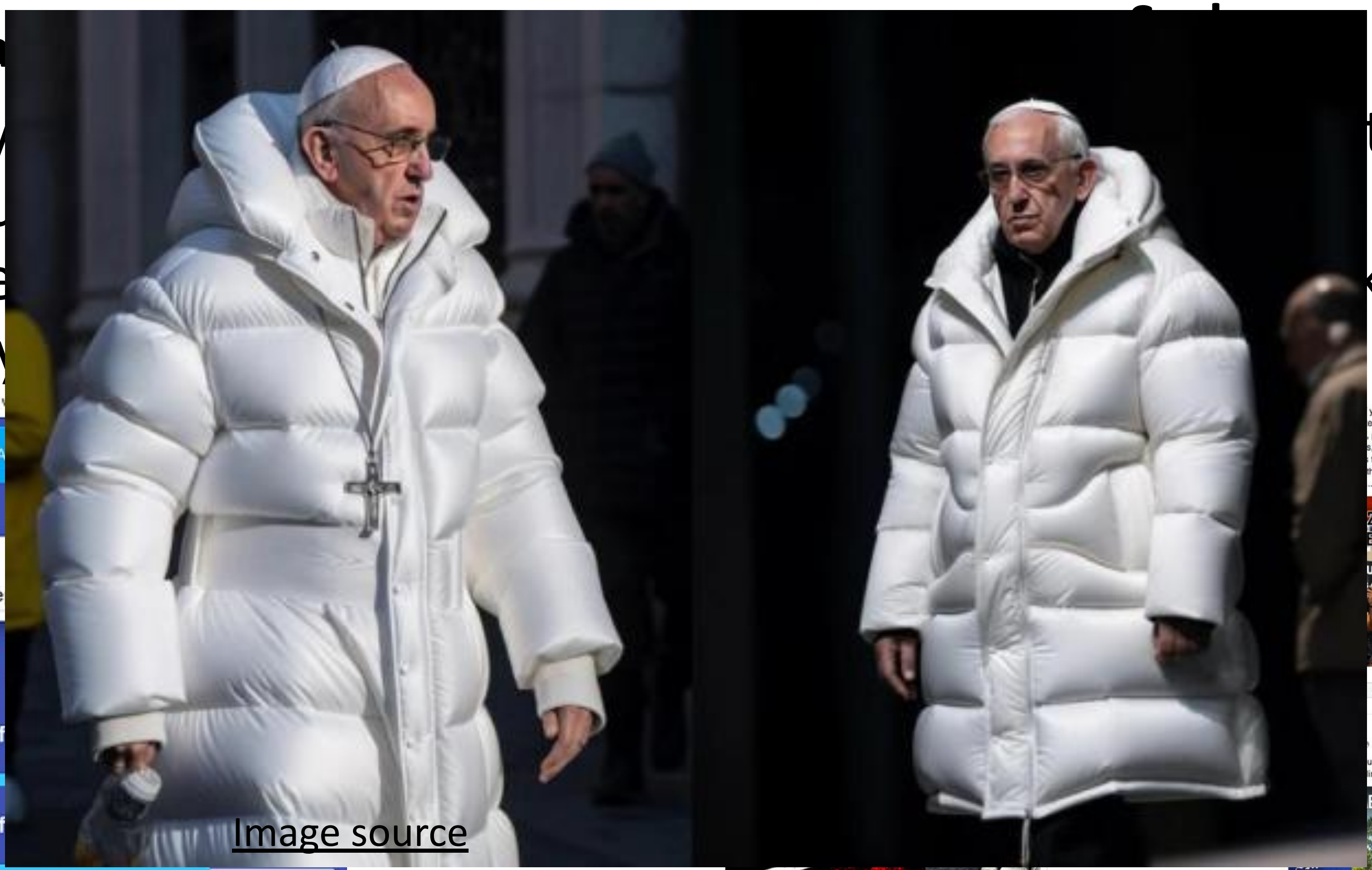


Image source

FOLLOW VETERANS US
IF YOU KNOW THE DIFFERENCE

Born Libe
Sponsored

american.veterans Killary Clinton will never understand what it feels like to lose the person you love for the sake of your country.

Secured Borders
Sponsored

Every man should stand for our borders! Join!

NO INVADERS ALLOWED

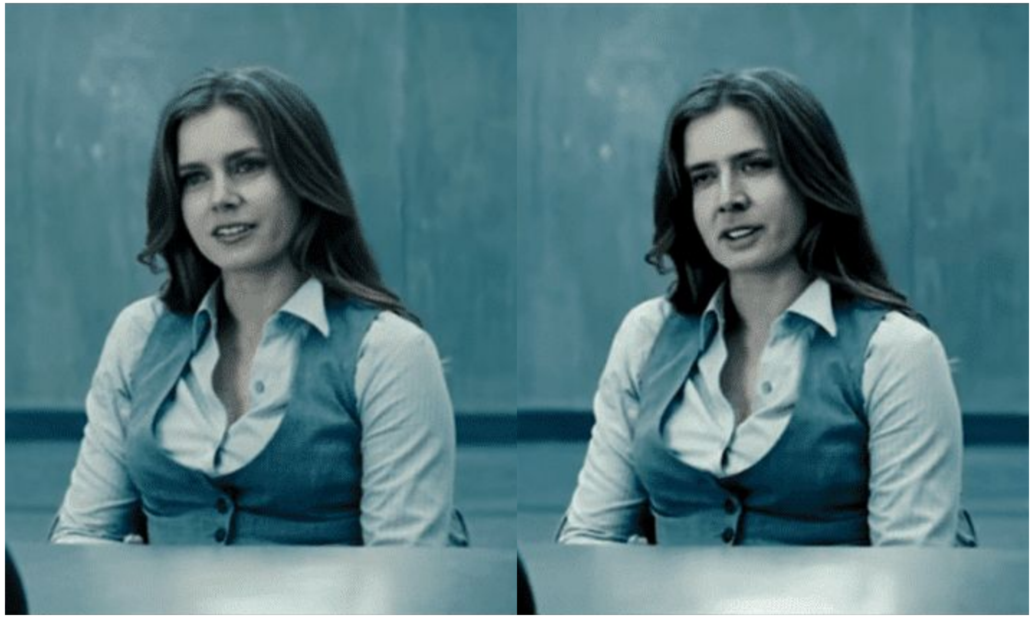
Secured Borders
News & Media Website
135,301 people like this.

.american.made
Sponsored

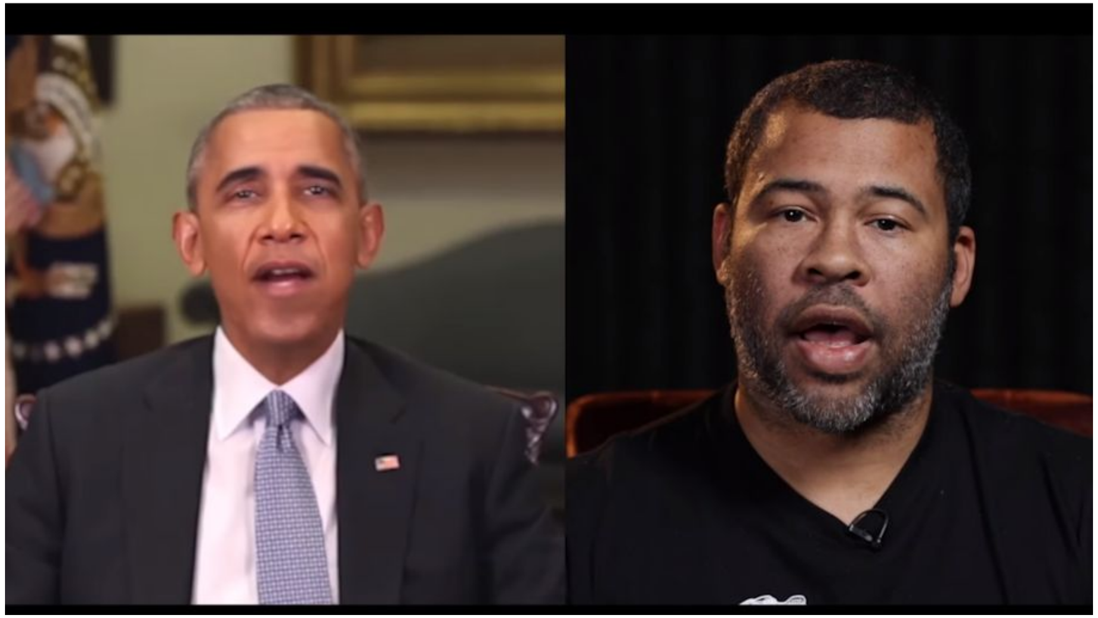
How are deepfakes made?

Two main “flavors” of deepfake

Face swap



Video puppetry



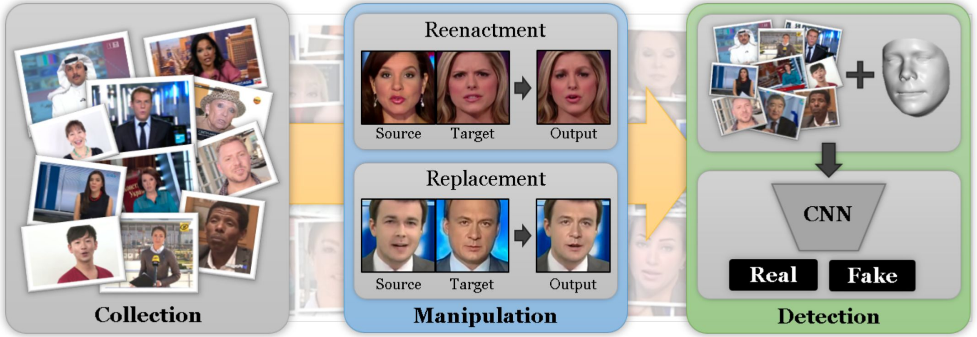
Alan Zucconi’s

Can deepfakes be stopped?

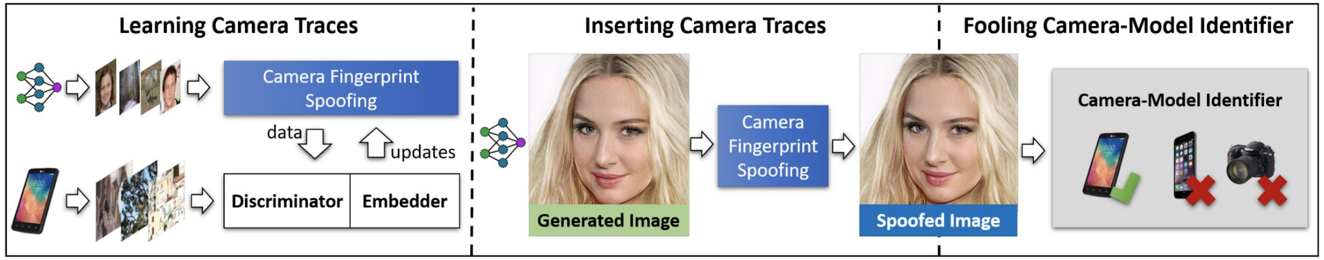
Detecting deepfakes

- **Deep learning**

- “Fighting fire with fire”



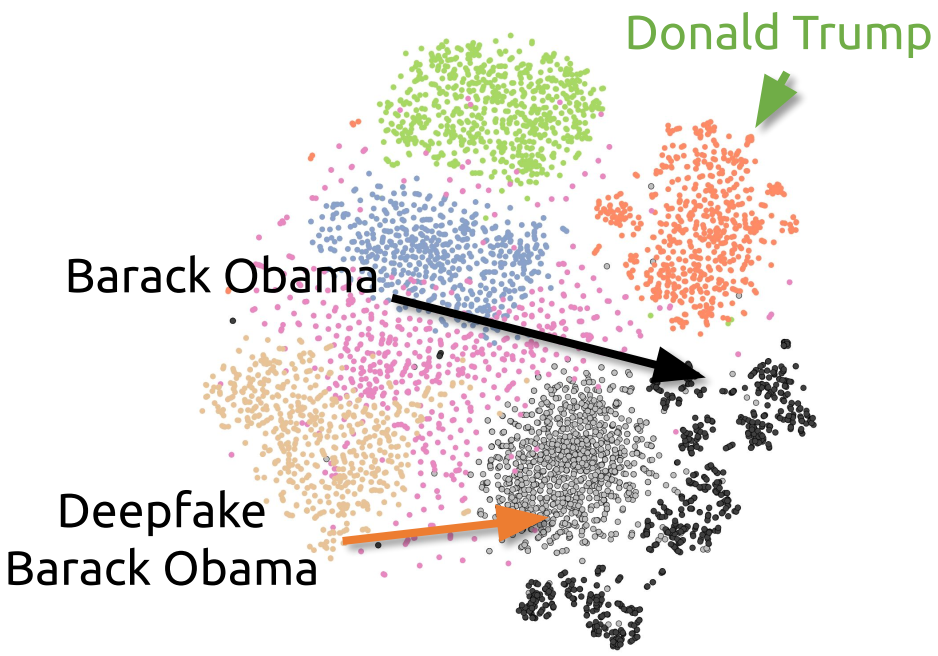
- ...but an adversary can train a model to fool your detector



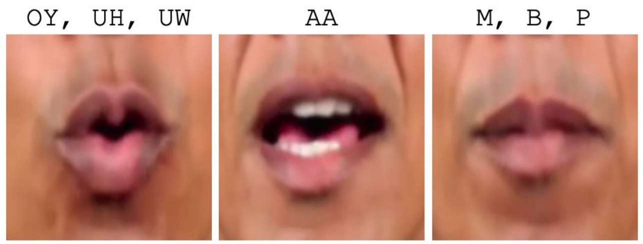
https://arxiv.org/pdf/1911.12069.pdf

Detecting deepfakes

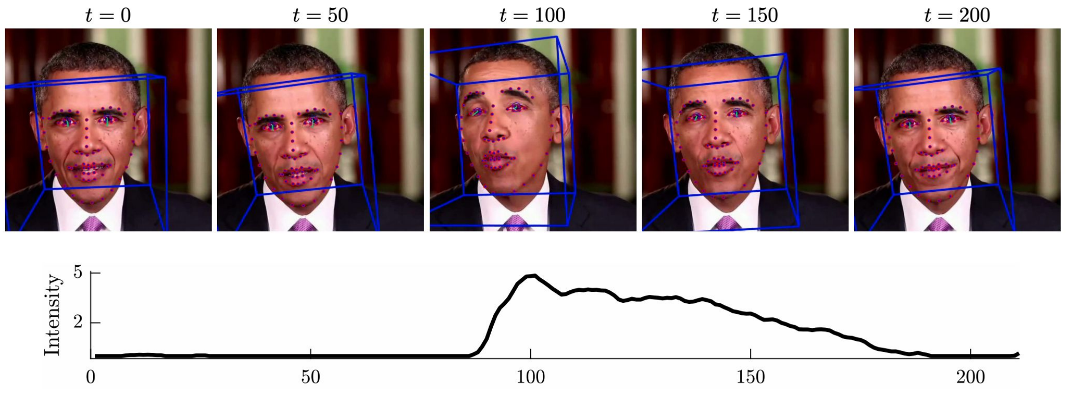
- Deep learning
- **“Classic” computer vision**



- Find inconsistencies between movements of lips and sounds

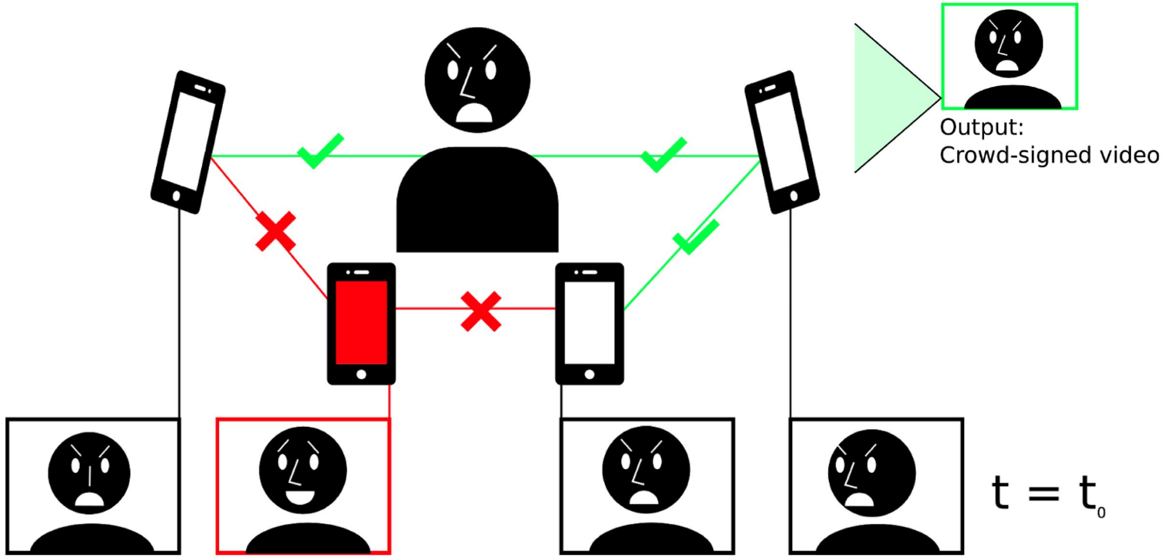


- Compute a “fingerprint” for a person based on how facial features tend to move over time



Detecting deepfakes

- Deep learning
- "Classic" computer vision
- **Social verification**



Parting thoughts

“What should I do about all this?”

- *If you’re working in ML/CV research:*
  - Think critically about, and articulate, the potential real-world impacts of your work (some conferences require this now)
  - Consider contributing to detection efforts if you also work on synthesis problems
- *If you’re working on user-facing products & services:*
  - Be vigilant for fake content on your platform
  - Initiate (and sustain) serious conversations with your coworkers and employers about how to responsibly take action
- *If you’re working in the government / non-profit sector:*
  - Help educate your less-technical colleagues about how deepfakes work
  - Support (or start!) movements to draft meaningful legislation

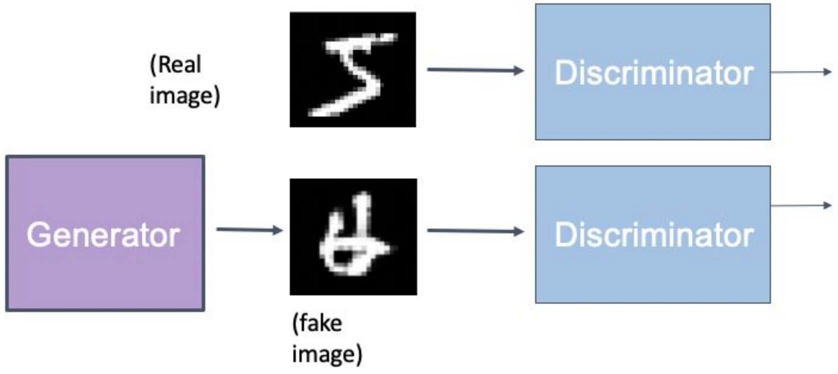
Recap

Generative Adversarial Networks (GANs)

- Architecture
- GAN Loss + Training
- Solving problem w/ GANs □ WGANs

Deepfakes

- What are deepfakes?
- Why are they a problem?
- How to detect deepfakes?



This Photo by Unknown Author is licensed under CC BY-NC-ND