Deep Learning for ECE EECE-580G

Introduction to Generative Adversarial Networks (GANs)

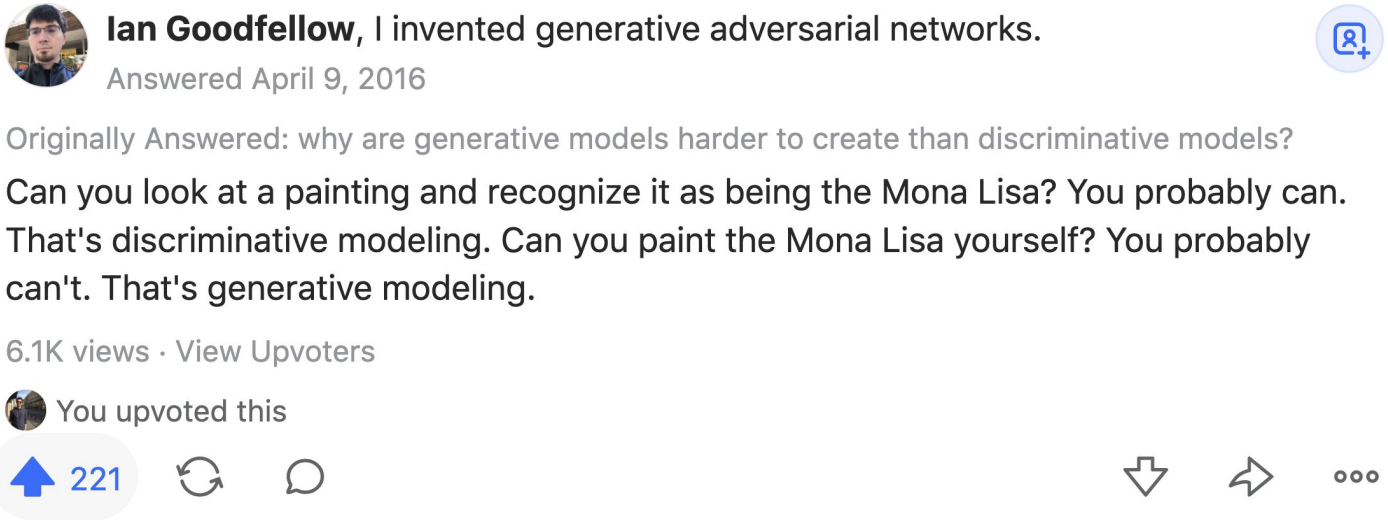
Introduction

Discriminative VS Generative

- Learn $p(y|x)$
- Assign labels to data
- Standard supervised learning

- Learn $p(x)$ The data distribution
- Do not need labels
- Or $p(x|y)$ Conditional generative model
- Learning data distributions is much more difficult

Discriminative VS Generative



Ian Goodfellow, I invented generative adversarial networks.

Answered April 9, 2016

Originally Answered: why are generative models harder to create than discriminative models?

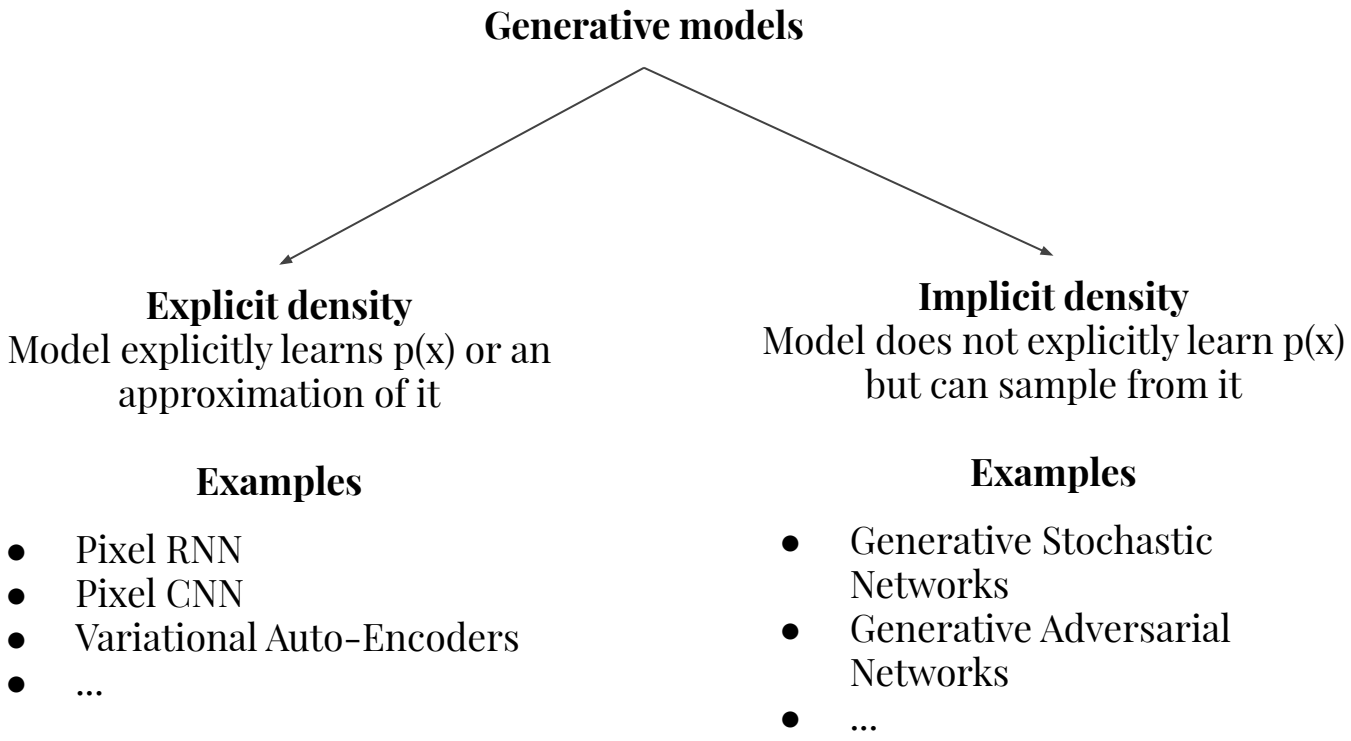
Can you look at a painting and recognize it as being the Mona Lisa? You probably can. That's discriminative modeling. Can you paint the Mona Lisa yourself? You probably can't. That's generative modeling.

6.1K views · View Upvoters

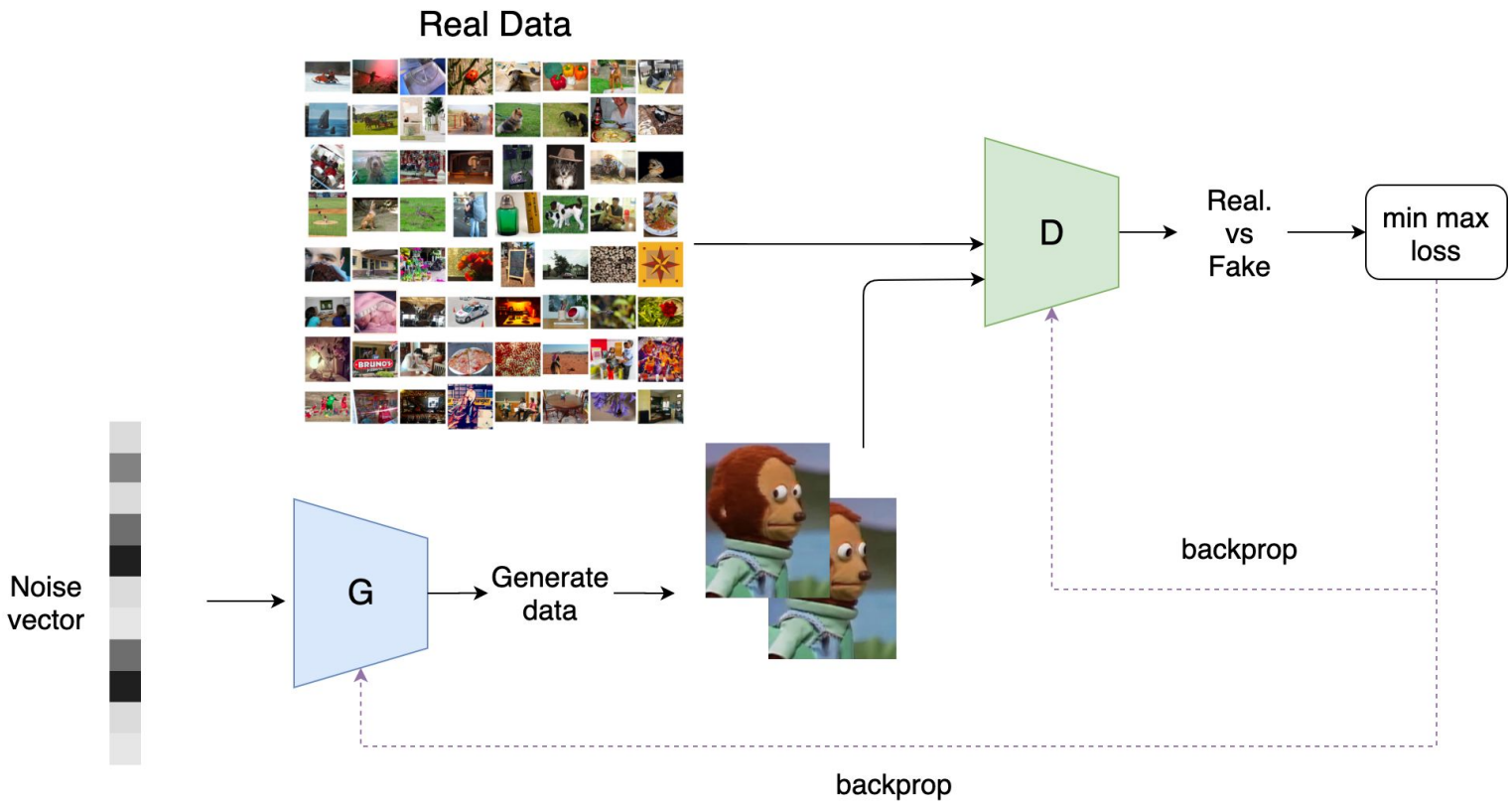
You upvoted this

221

Discriminative VS Generative



GANs

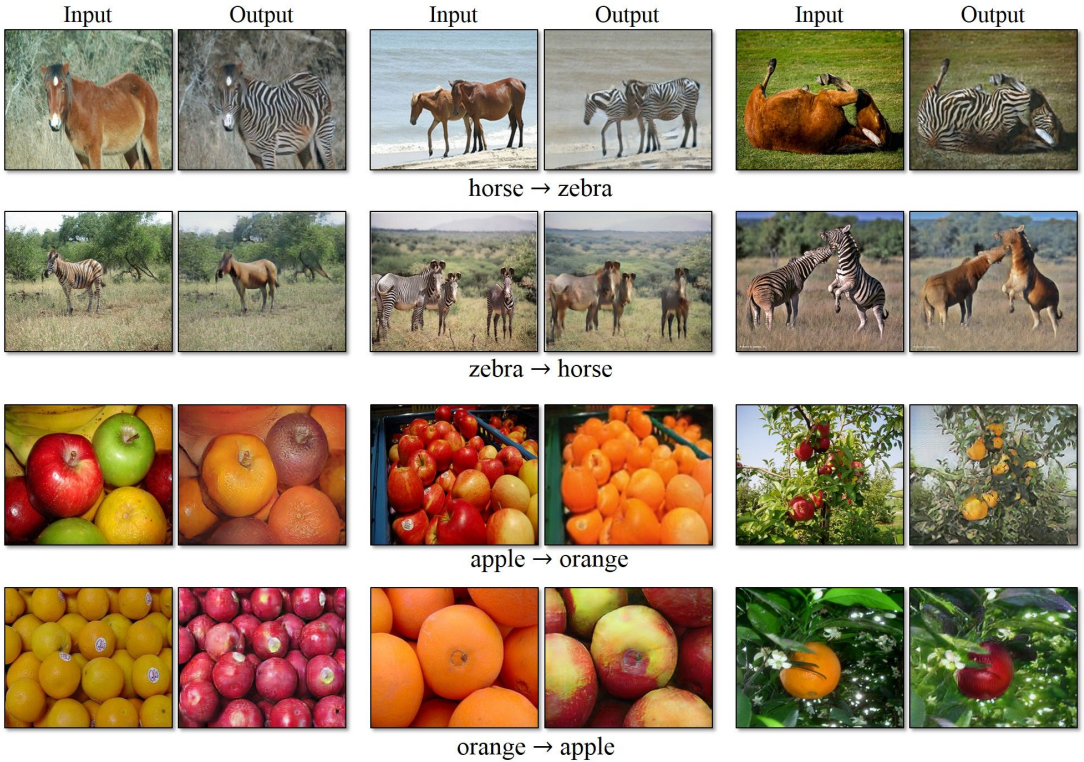

Real Data
D
Real.
vs
Fake
min max
loss
Noise
vector
G
Generate
data
backprop
backprop

GANs

Applications

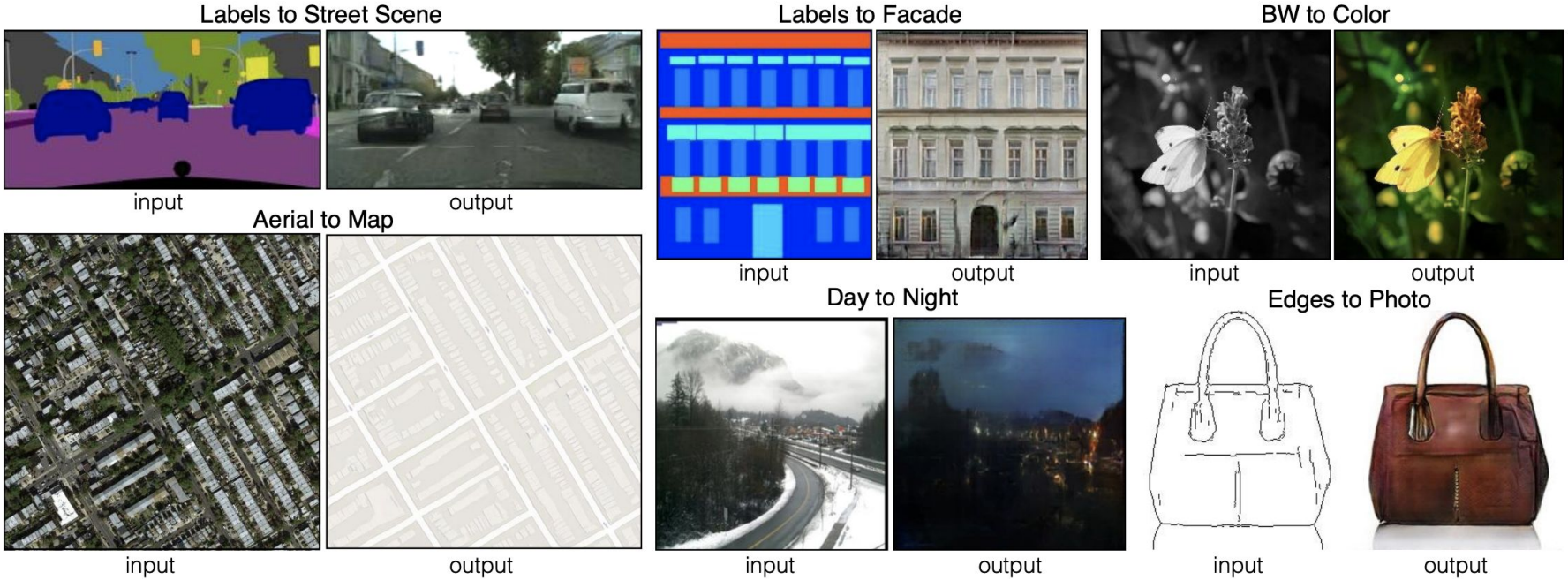
Evolution of GANs

Image to image translation (unpaired)



Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.

Image to image translation (paired)



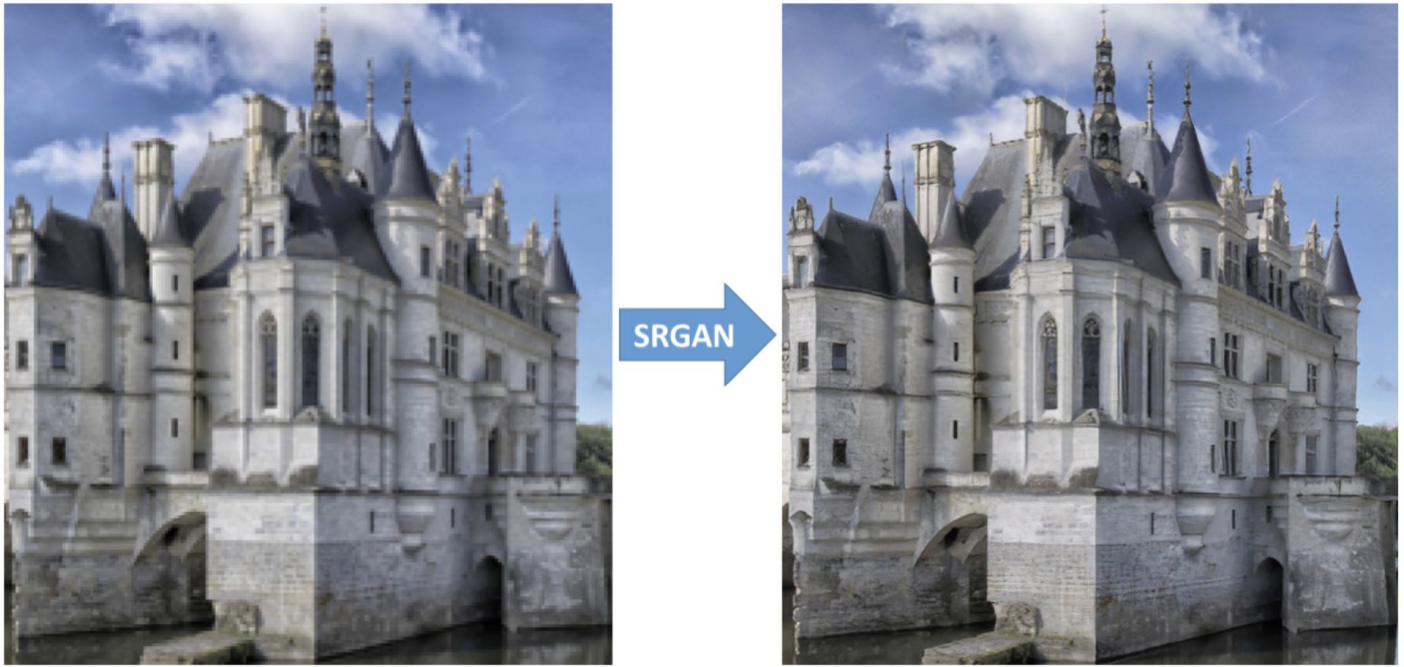
Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

Dataset augmentation

Figure 6: Samples generated by our BigGAN model at 512×512 resolution.

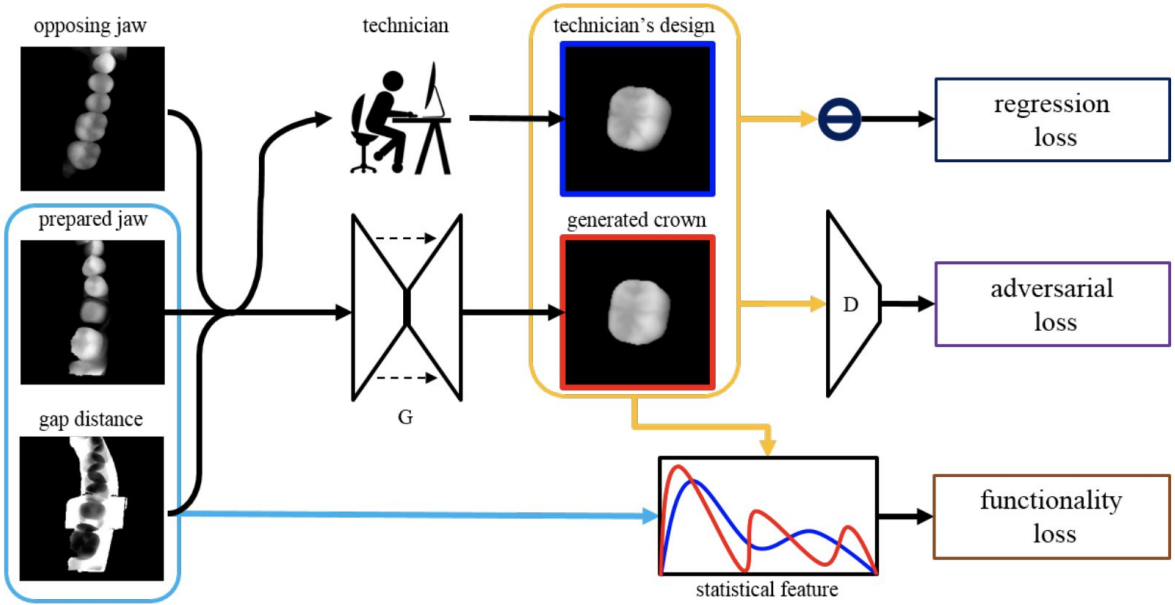
Brock, Andrew, Jeff Donahue, and Karen Simonyan. "Large scale gan training for high fidelity natural image synthesis." arXiv preprint arXiv:1809.11096 (2018).

Super resolution



Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

GANufacturing



Hwang, Jyh-Jing, et al. "Learning beyond human expertise with generative models for dental restorations." arXiv preprint arXiv:1804.00064 (2018).

Nvidia Maxine (Face Alignment)

https://bdtechtalks.com/2020/10/19/nvidia-maxine-ai-video-conferencing/
NVIDIA Maxine

Climate change awareness

Zhou, Sharon, et al. "Establishing an evaluation metric to quantify climate change image realism." Machine Learning: Science and Technology 1.2 (2020): 025005.

Steganography

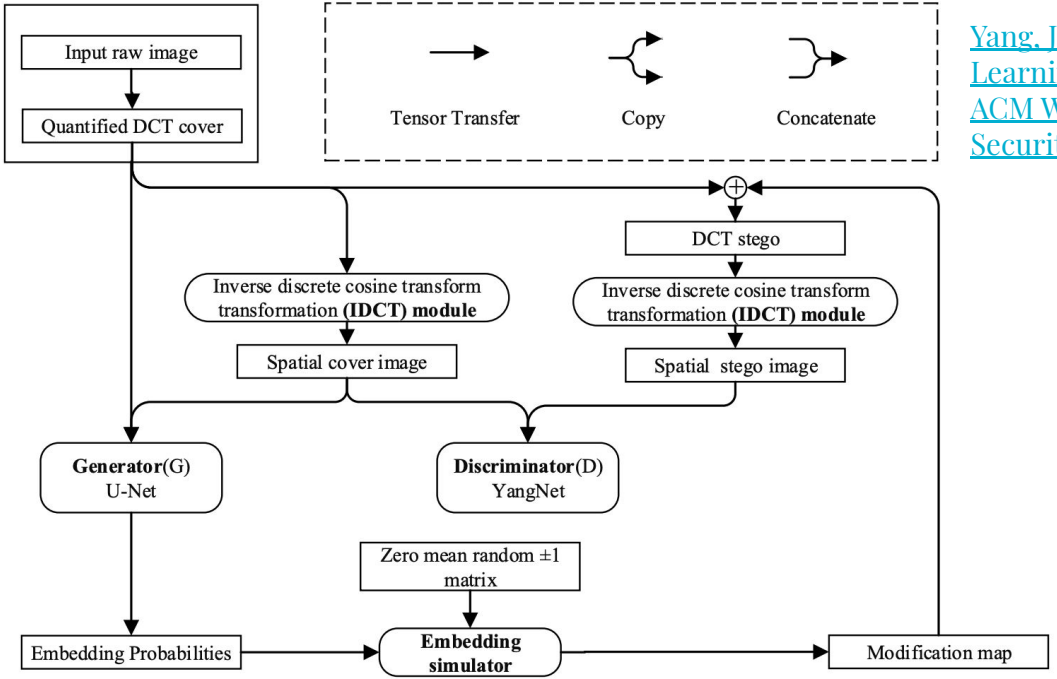
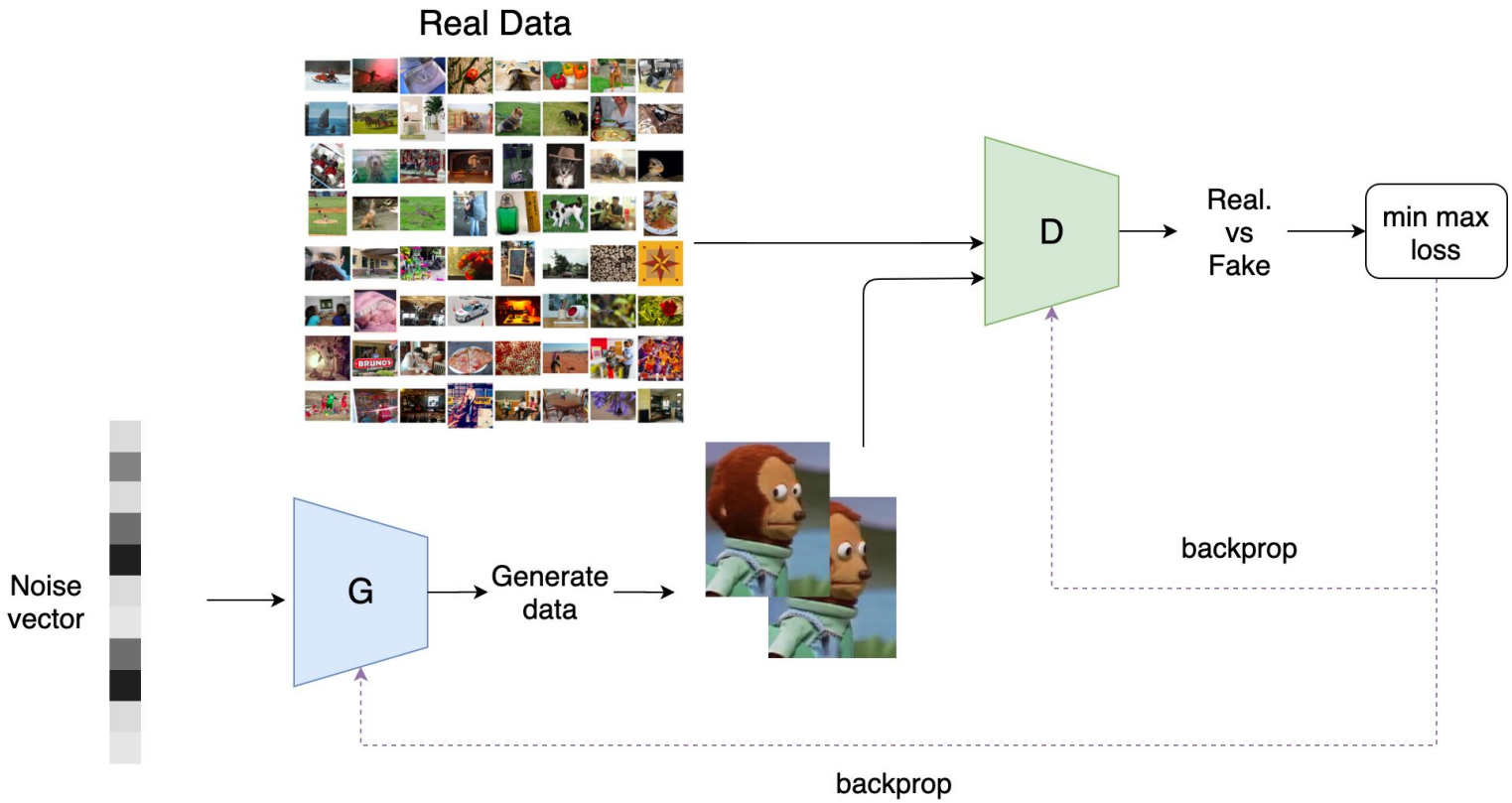


Figure 1: Steganographic architecture of the proposed JS-GAN.

Yang, Jianhua, et al. "Towards Automatic Embedding Cost Learning for JPEG Steganography." Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. 2019.

GANs

Concept


Real Data
D
Real.
vs
Fake
min max
loss
Noise
vector
G
Generate
data
backprop
backprop

Algorithm

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

  for k steps **do**

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
  - Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

  end for

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

D and G are MLPs

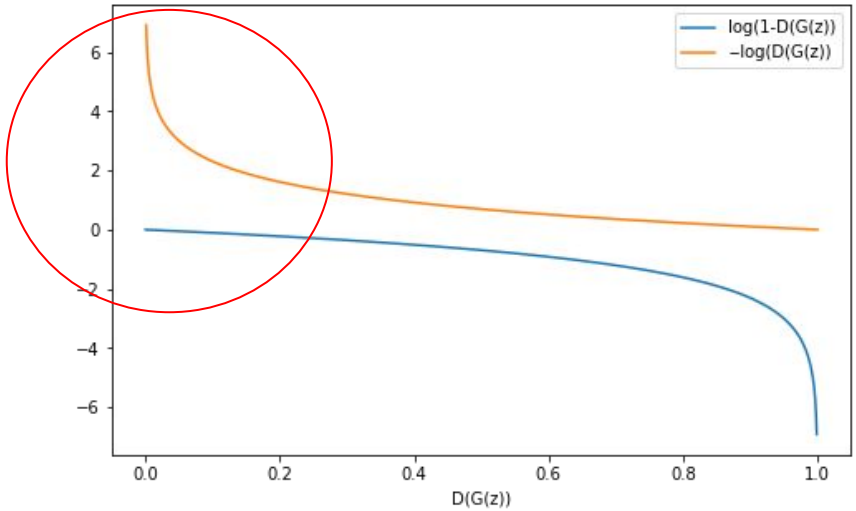
Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

Optimality

- Global minimum of the min-max loss occurs when $p_g = p_{data}$
- Proof in Section 4.1 of [Goodfellow, Ian, et al. "Generative adversarial nets." NeurIPS 2014]
- Does not tell us if the algorithm reaches the minimum in practice 😕

Saturation

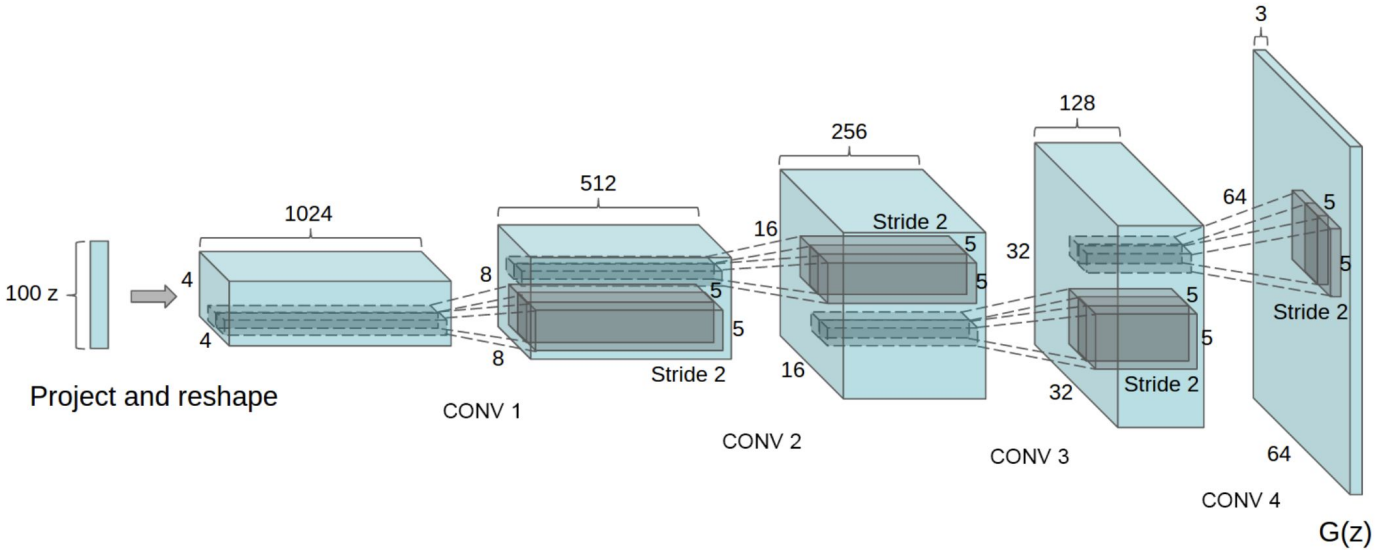
- At early stages: generator is bad, discriminator can easily output o for G(z)
- Gradient given to the generator is weak (blue curve)
- Use a different loss for generator
- $-log(D(G(z)))$
- Called Non-Staturating GANs (NS-GAN)
- No longer zero-sum different functions to optimize for gen and disc
- Both loss functions can be written using BCE loss



DCGANs

Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

DC-GANs



Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

Tricks for more stable training

Some of these tricks have been challenged in recent advances

Architecture guidelines for stable Deep Convolutional GANs

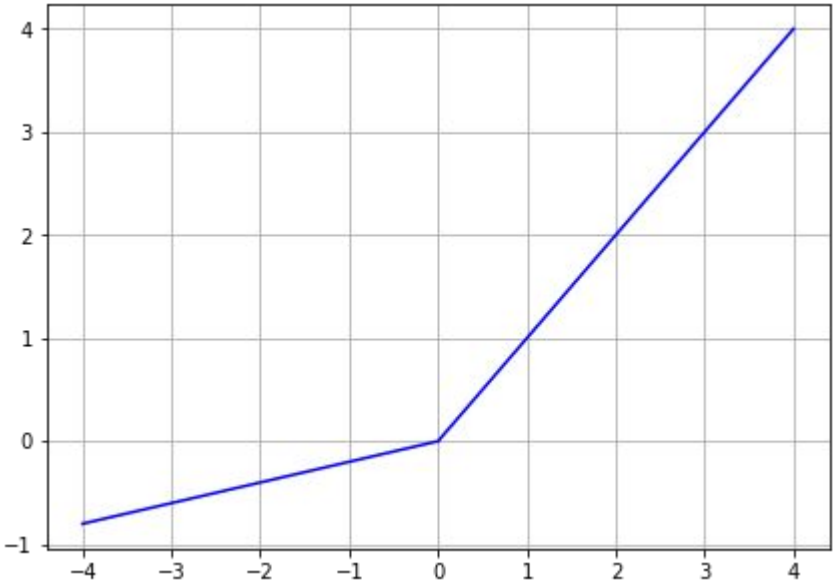
- Replace any pooling layers with strided convolutions (discriminator) and fractional-strided convolutions (generator).
- Use batchnorm in both the generator and the discriminator.
- Remove fully connected hidden layers for deeper architectures.
- Use ReLU activation in generator for all layers except for the output, which uses Tanh.
- Use LeakyReLU activation in the discriminator for all layers.

Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

Leaky ReLU

$$\varphi(z) = \max\{az, z\}$$

$$0 < a < 1$$



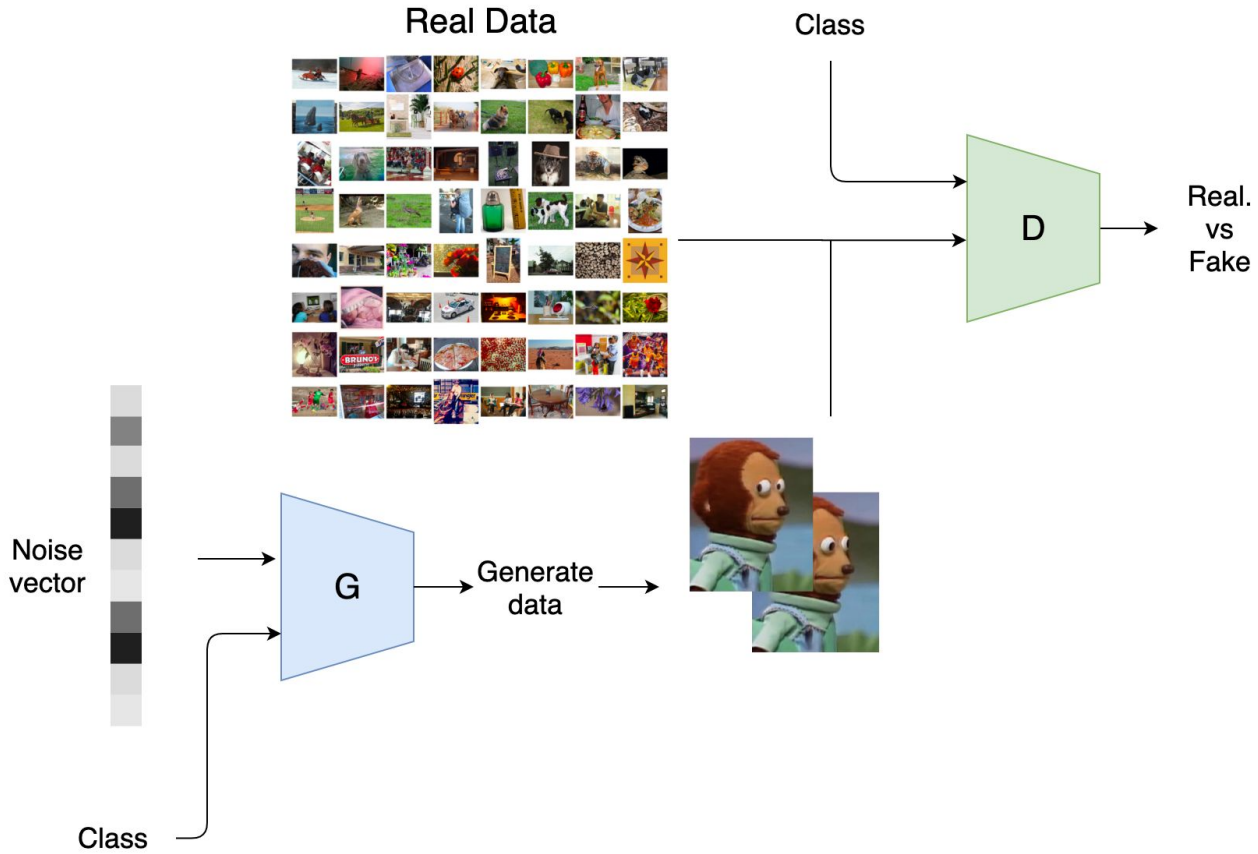
Conditional GANs

Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." arXiv preprint arXiv:1411.1784 (2014).

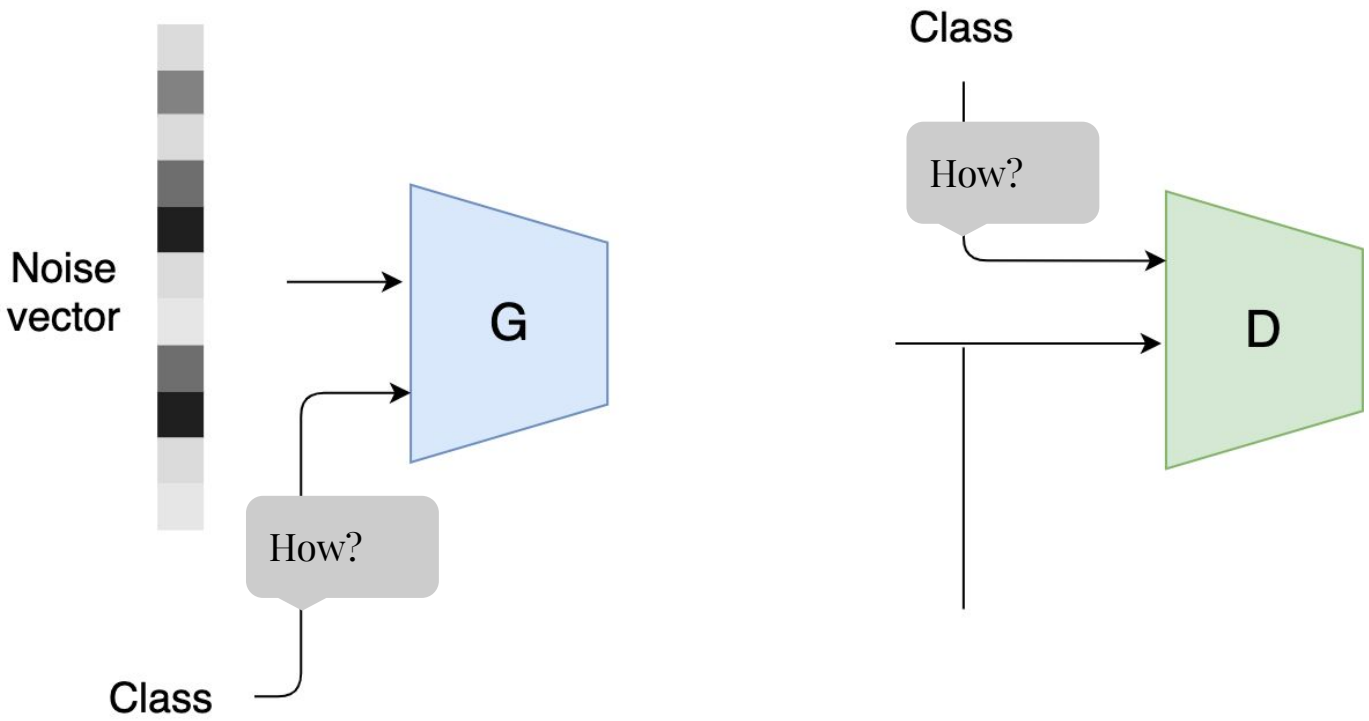
Conditional Generative Models

- We want the Generator not only to produce a new image, but a new image of a desired class (or label)
- i.e. learn $p(x|y)$
- First introduced in Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." arXiv preprint arXiv:1411.1784 (2014).
- But now widely used
- **Using labels when available has proven to produce better GANs even if we are not interested in conditional generation**

Concept

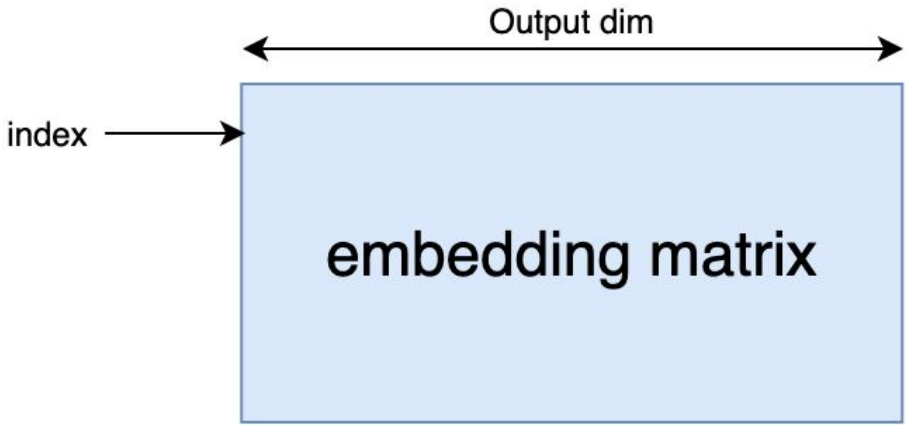

Real Data
Class
D
Real.
vs
Fake
Noise
vector
G
Generate
data
Class

Concept



The embedding layer

- Solution #1 = one-hot encoding + reshaping
- Solution #2 = learnable embedding layer



Wasserstein GAN / Gradient Penalty

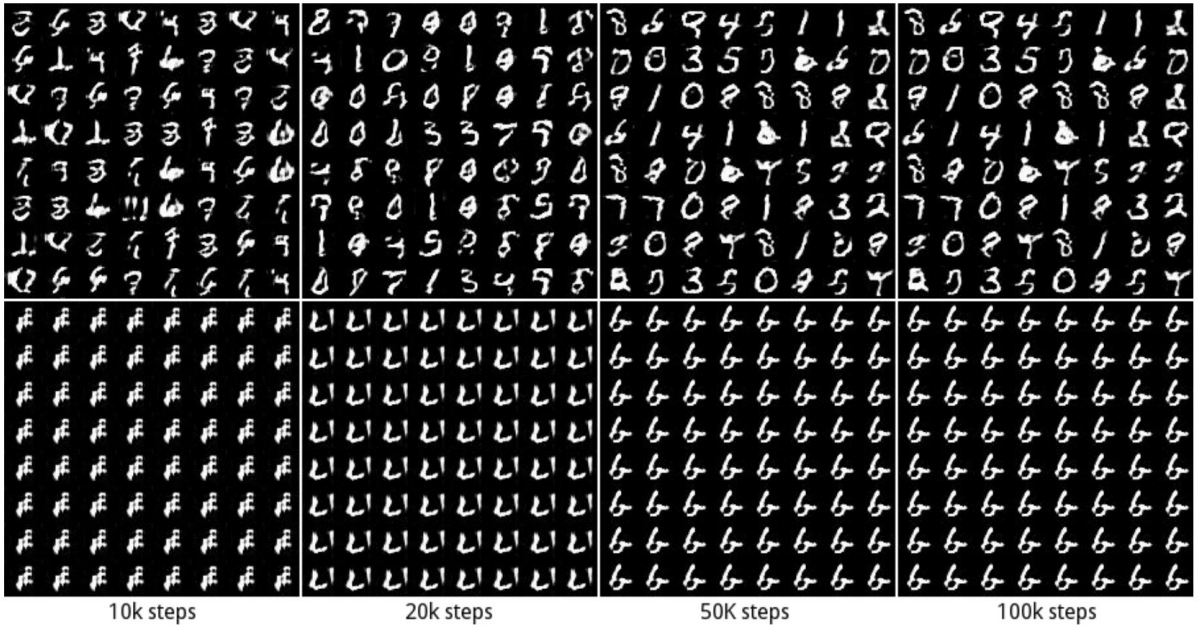
Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein gan." arXiv preprint arXiv:1701.07875 (2017).

Gulrajani, Ishaan, et al. "Improved training of wasserstein gans." Advances in neural information processing systems. 2017.

Shortcomings of BCE (mode collapse)

- Generator produces an especially plausible output but same for all seeds (or small set)
- The discriminator's best strategy is to learn to always reject that output
- Suppose discriminator gets stuck in a local minimum and doesn't find this best strategy
- The next generator iteration will only produce those plausible outputs
- Each iteration of generator over-optimizes for a particular discriminator, and the discriminator never manages to learn its way out of the trap

Shortcomings of BCE (mode collapse)



Metz, Luke, et al. "Unrolled generative adversarial networks." arXiv preprint arXiv:1611.02163 (2016).

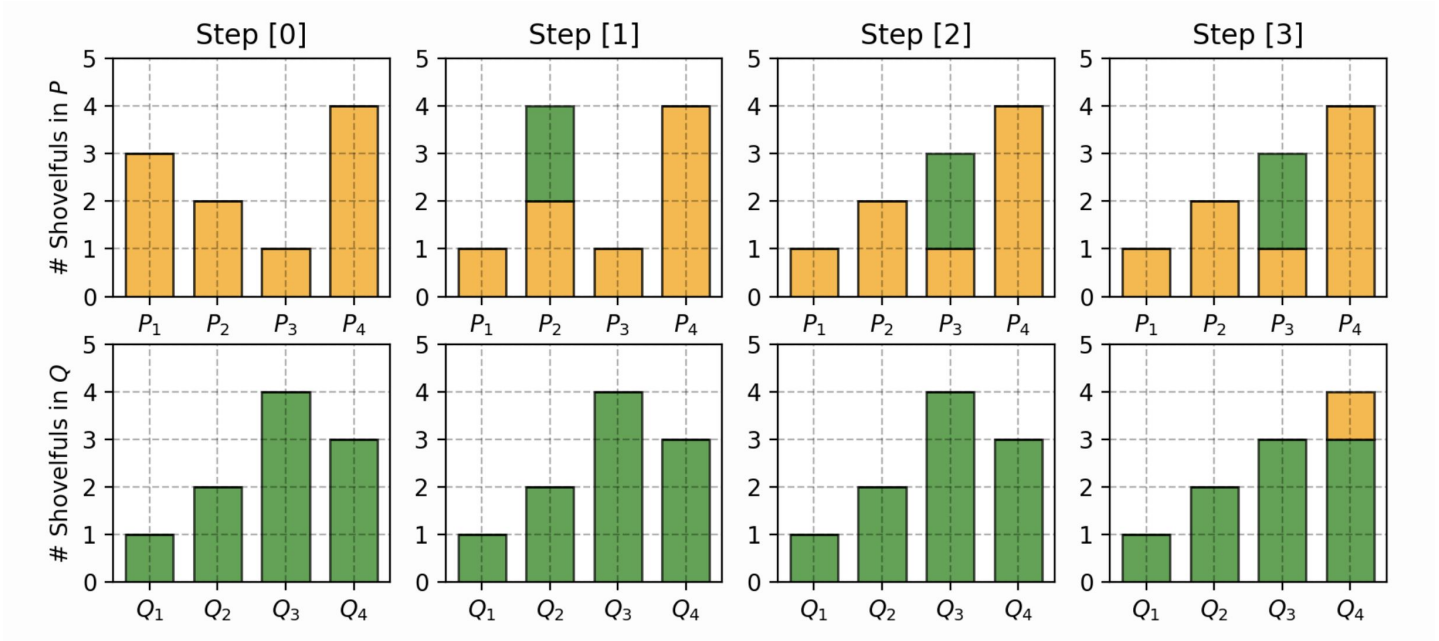
Shortcomings of BCE (mode collapse)

- Generator produces an especially plausible output but same for all seeds (or small set)
- The discriminator's best strategy is to learn to always reject that output
- Suppose discriminator gets stuck in a local minimum and doesn't find this best strategy
- The next generator iteration will only produce those plausible outputs
- Each iteration of generator over-optimizes for a particular discriminator, and the discriminator never manages to learn its way out of the trap
- **How can we fix this?**
- **(A)** Make sure discriminator trains well?
- => Can lead to discriminator too good => vanishing gradients
- **(B)** Change objective function?

Wasserstein distance (earth mover's distance)

https://youtu.be/zo46TEp6FB8

Wasserstein-1 distance (earth mover's distance)



https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html

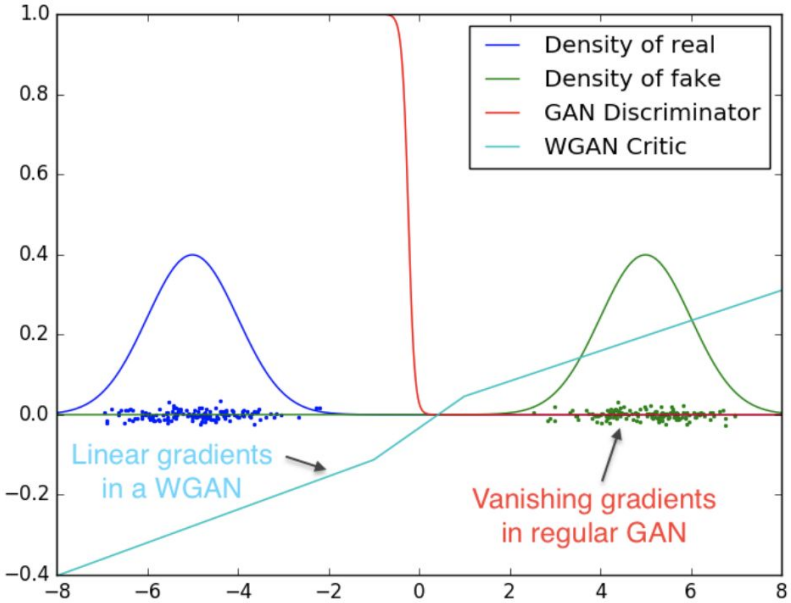
Kantorovich-Rubinstein duality

$$\text{EMD} = \sup_{\|f\|_{L \leq 1}} \mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{x \sim p_g} f(x)$$

- (discriminator) Critic maximizes the expression (approximates EMD)
- Generator minimizes EMD

nobelprize.org

W-GAN



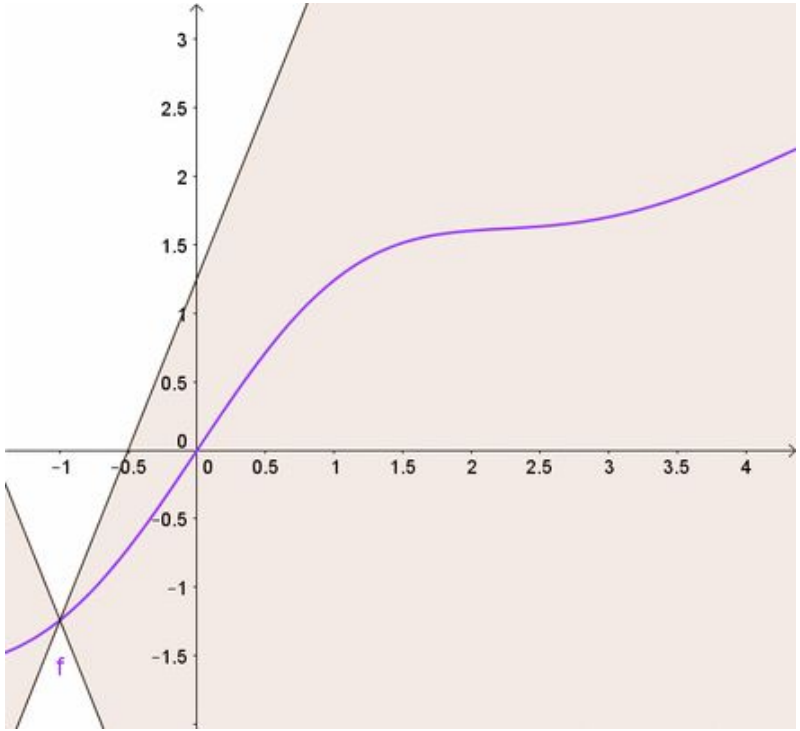
Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein gan." arXiv preprint arXiv:1701.07875 (2017).

Lipschitz continuity - Gradient Penalty

$$\text{EMD} = \sup_{\|f\|_{L \leq 1}} \mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{x \sim p_g} f(x)$$

- Critic has to be (at-most) 1-Lipschitz continuous
- (real valued) critic f is K-Lipschitz continuous iif $|f(x_1) - f(x_2)| \leq K\|x_1 - x_2\|_2$

Lipschitz continuity - Gradient Penalty



https://en.wikipedia.org/wiki/Lipschitz_continuity

Lipschitz continuity - Gradient Penalty

$$\text{EMD} = \sup_{\|f\|_{L \leq 1}} \mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{x \sim p_g} f(x)$$

- **Critic has to be (at-most) 1-Lipschitz continuous**
- (real valued) critic f is K-Lipschitz continuous iif $|f(x_1) - f(x_2)| \leq K \|x_1 - x_2\|_2$
- If f is **differentiable** (it is) f is K-Lipschitz continuous iif $\|\nabla f(x)\|_2 \leq K$

wGAN-GP

$$\min \max \mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{x \sim p_g} f(x) + \lambda R(f)$$

$$R(f) = (\|\nabla f(x)\|_2 - 1)^2$$

wGAN-GP

$$\min \max \mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{x \sim p_g} f(x) + \lambda R(f)$$

$$R(f) = \left(\|\nabla f(x)\|_2 - 1 \right)^2$$

Algorithm 1 WGAN with gradient penalty. We use default values of $\lambda = 10$, $n_{\text{critic}} = 5$, $\alpha = 0.0001$, $\beta_1 = 0$, $\beta_2 = 0.9$.

Require: The gradient penalty coefficient λ, the number of critic iterations per generator iteration n_{critic}, the batch size m, Adam hyperparameters α, β_1, β_2.
Require: initial critic parameters w_0, initial generator parameters θ_0.
while θ has not converged **do**
  for $t = 1, ..., n_{\text{critic}}$ **do**
    for $i = 1, ..., m$ **do**
      Sample real data $\boldsymbol{x} \sim \mathbb{P}_r$, latent variable $\boldsymbol{z} \sim p(\boldsymbol{z})$, a random number $\epsilon \sim U[0, 1]$.
      $\tilde{\boldsymbol{x}} \leftarrow G_\theta(\boldsymbol{z})$
      $\hat{\boldsymbol{x}} \leftarrow \epsilon \boldsymbol{x} + (1 - \epsilon)\tilde{\boldsymbol{x}}$
      $L^{(i)} \leftarrow D_w(\tilde{\boldsymbol{x}}) - D_w(\boldsymbol{x}) + \lambda(\|\nabla_{\hat{\boldsymbol{x}}} D_w(\hat{\boldsymbol{x}})\|_2 - 1)^2$
    end for
    $w \leftarrow \text{Adam}(\nabla_w \frac{1}{m} \sum_{i=1}^m L^{(i)}, w, \alpha, \beta_1, \beta_2)$
  end for
  Sample a batch of latent variables $\{\boldsymbol{z}^{(i)}\}_{i=1}^m \sim p(\boldsymbol{z})$.
  $\theta \leftarrow \text{Adam}(\nabla_\theta \frac{1}{m} \sum_{i=1}^m -D_w(G_\theta(\boldsymbol{z})), \theta, \alpha, \beta_1, \beta_2)$
end while

Gulrajani, Ishaan, et al. "Improved training of wasserstein gans." Advances in neural information processing systems. 2017.

Other regularization techniques and GANs tricks

– – –

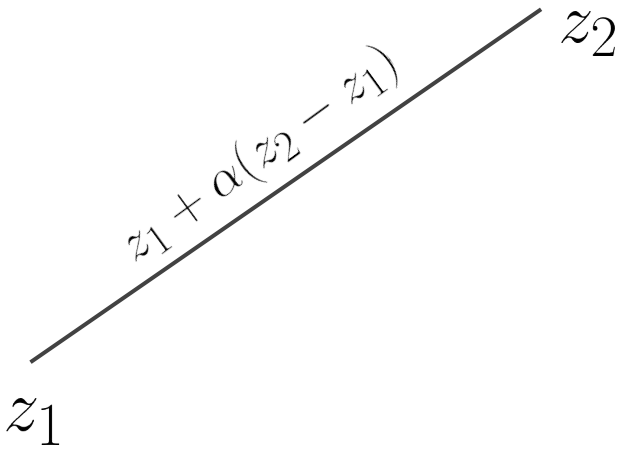
- Spectral normalization Miyato, Takeru, et al. "Spectral normalization for generative adversarial networks." arXiv preprint arXiv:1802.05957 (2018).
  Implemented here https://www.tensorflow.org/addons/api_docs/python/tfa/layers/SpectralNormalization
- Orthogonal regularization Saxe, Andrew M., James L. McClelland, and Surya Ganguli. "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks." arXiv preprint arXiv:1312.6120 (2013).
- Label smoothing (one sided)
- Weight averaging Salimans, Tim, et al. "Improved techniques for training gans." Advances in neural information processing systems. 2016. and Yazici, Yasin, et al. "The unusual effectiveness of averaging in GAN training." (2019).
- Other Batch Normalization layers (virtual BN, Layer normalization, conditional BN, ...)

Sampling GANs

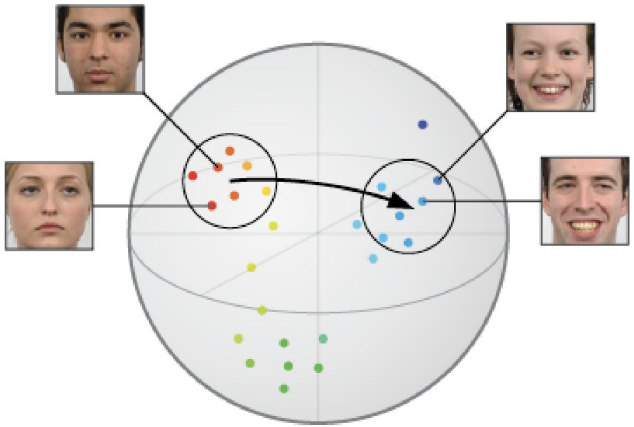
Interpolation

Interpolation

Linear interpolation

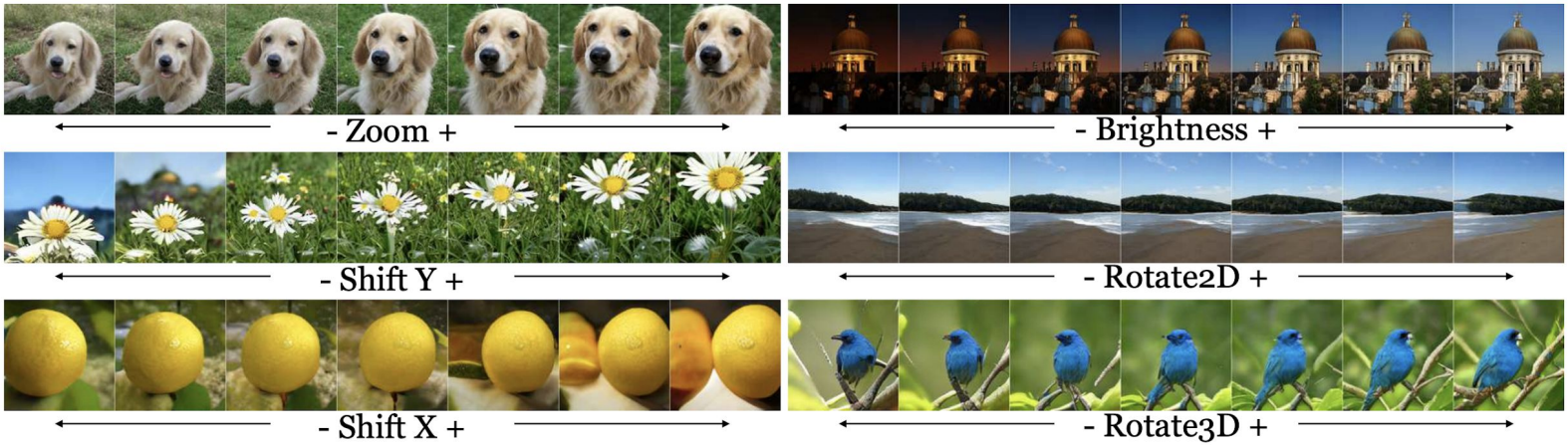


Spherical interpolation (Slerp)



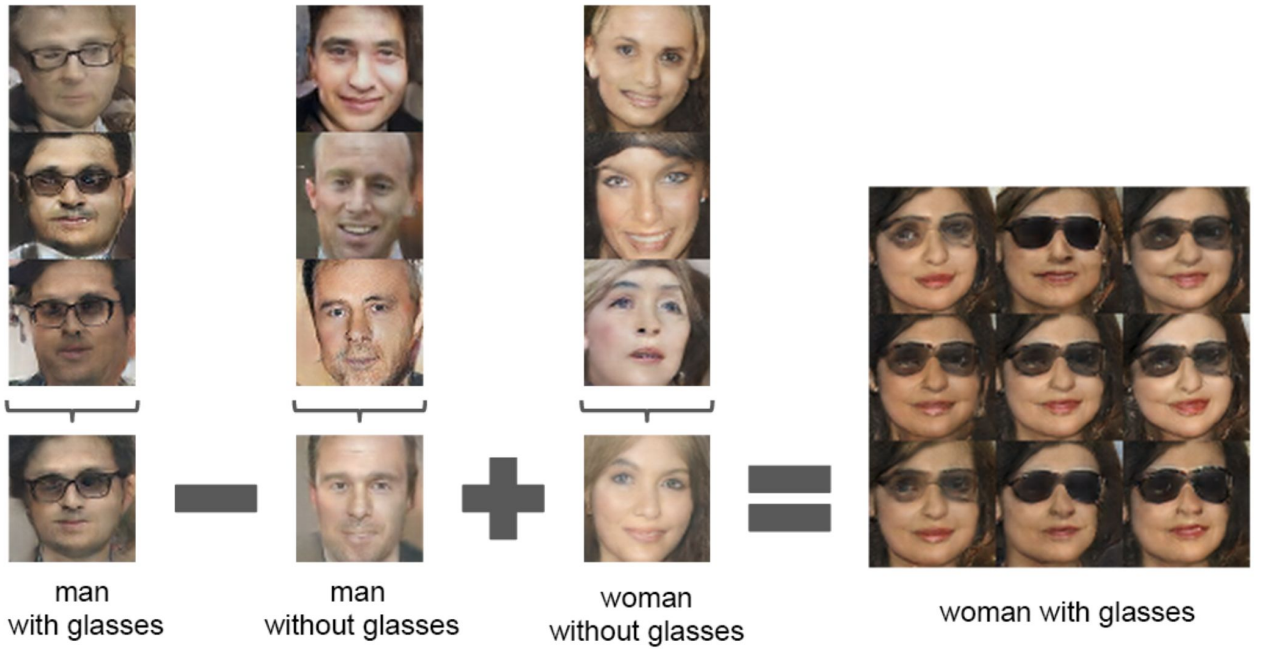
White, Tom. "Sampling generative networks." arXiv preprint arXiv:1609.04468 (2016).

Steering in the z space



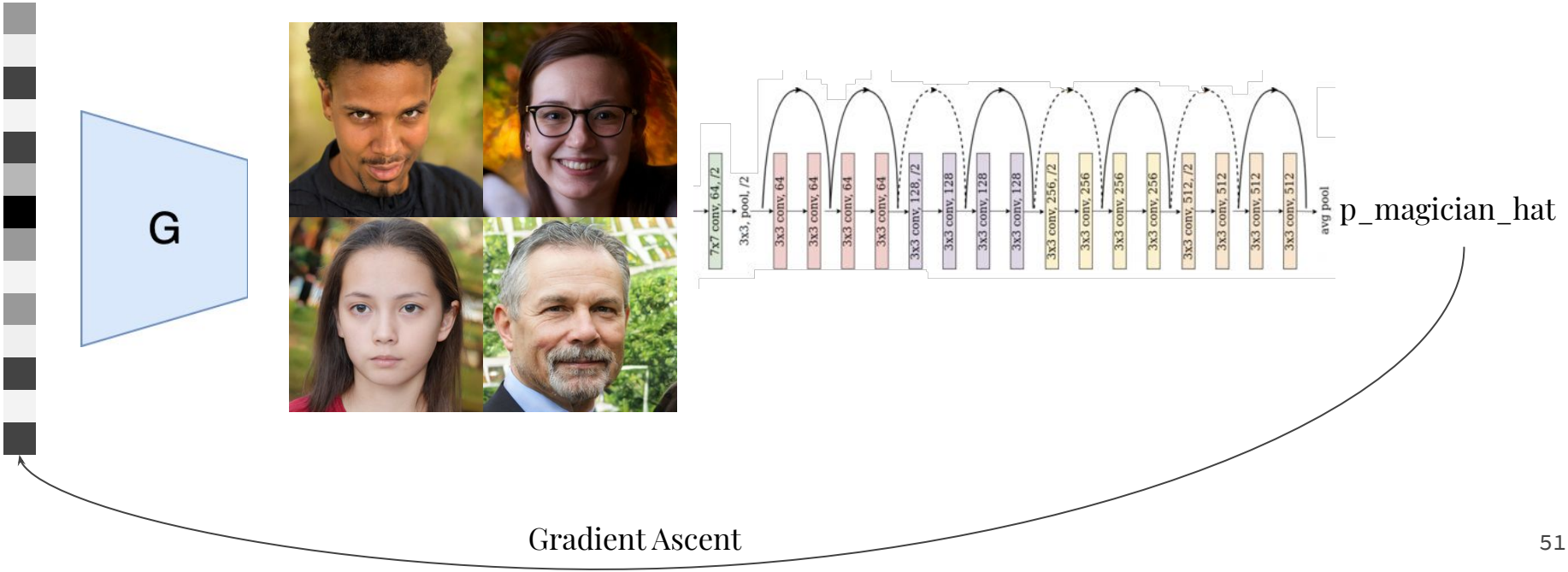
Jahanian, Ali, Lucy Chai, and Phillip Isola. "On the"steerability" of generative adversarial networks." arXiv preprint arXiv:1907.07171 (2019).

Vector arithmetic

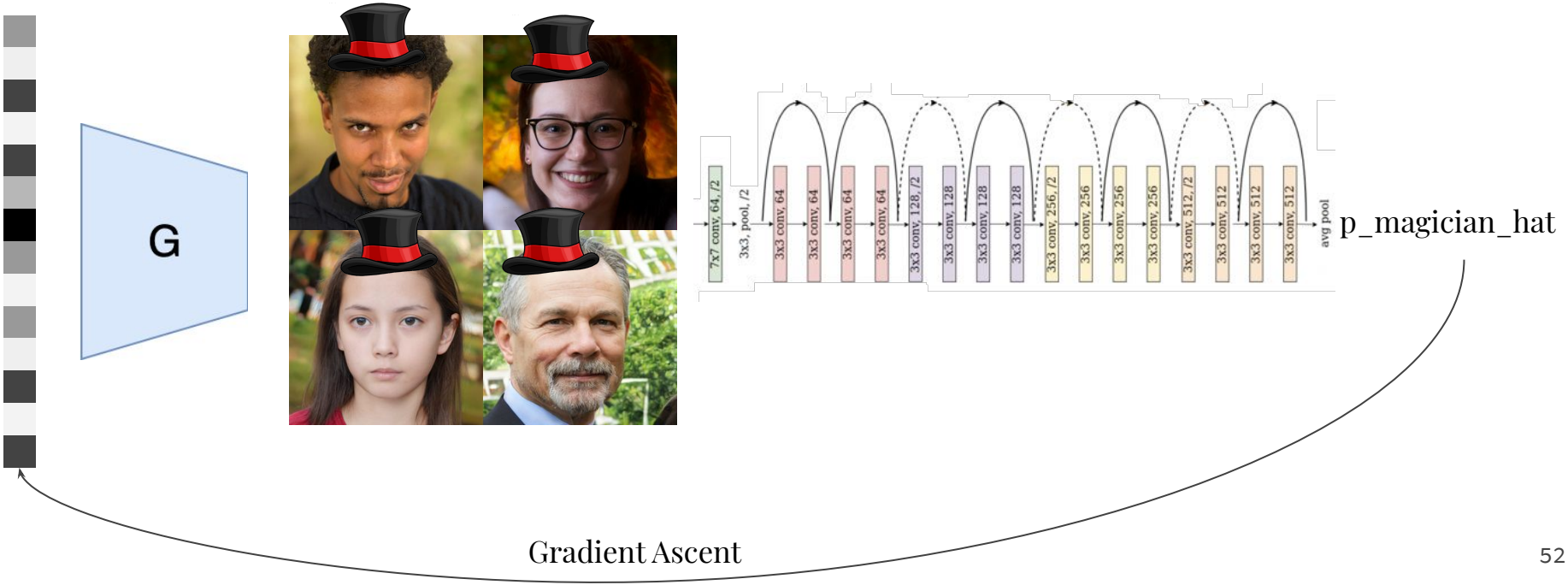


Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

Classifier gradients

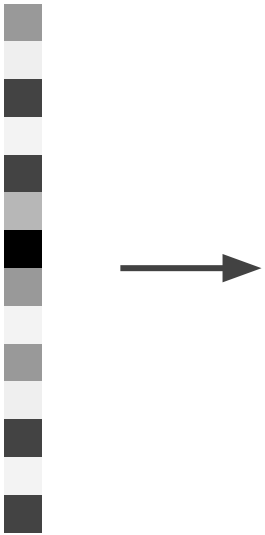


Classifier gradients



GAN Evaluation

How to evaluate GANs

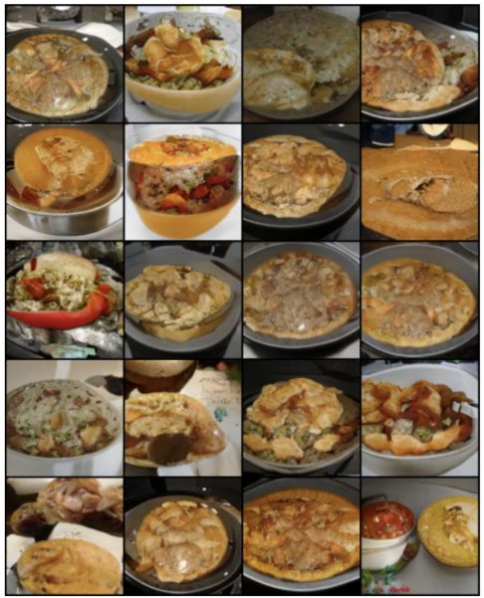


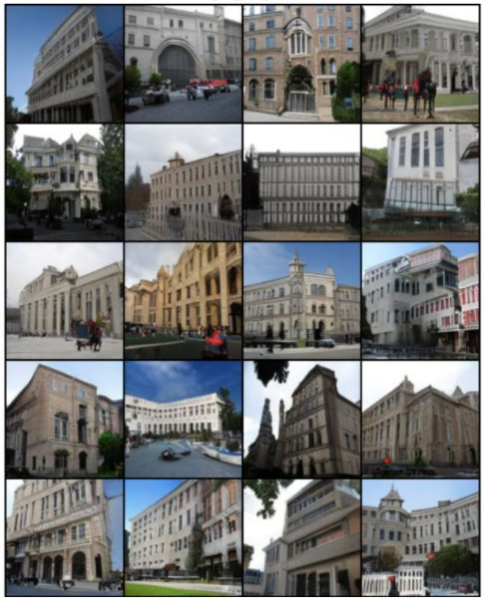
https://www.kaggle.com/c/generative-dog-images/

Fidelity

https://thispersondoesnotexist.com/

Diversity

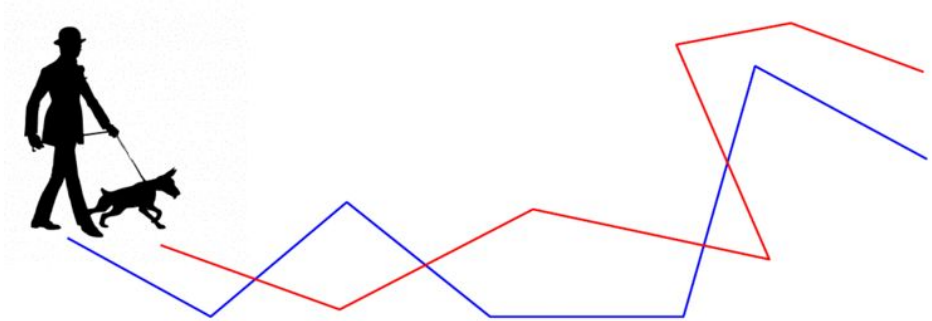




Liu, Steven, et al. "Diverse Image Generation via Self-Conditioned GANs." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

Frechet Inception Distance

- A person and a dog connected by a leash
- Each walking along a different curve from its starting point to its end point.
- Both can control their speed, but not allowed to backtrack.
- **The Fréchet distance between the two curves is the minimum length of a leash that is sufficient for traversing both curves in this manner.**
- Special case (p=2) of Wasserstein Distances



Credits: https://omrit.filtser.com/

Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems. 2017.

Frechet Inception Distance - Gaussian Case

- When the two “curves” are Gaussian pdf’s

$$d^2 = \|\mu_1 - \mu_2\|^2 + tr(\Sigma_1 + \Sigma_2 - 2\sqrt{\Sigma_1\Sigma_2})$$

- Use inception-v3 model to extract feature representations of real and generated images
- Approximate the distributions as multivariate Gaussian distributions
- Estimate the moments
- Compute FID

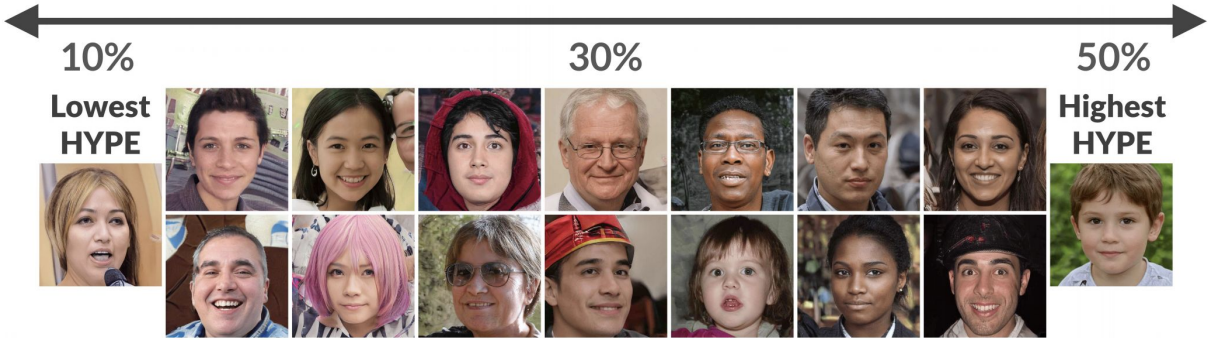
Frechet Inception Distance

- Need to use a large number of samples (typically 50,000 images)
- Usually more samples ↠ lower FID
- Pre-trained imagenet inception-v3 might not extract relevant features for some types of images (i.e. images very different from imagenet distribution)
- Only moments matching
- Slow to run if not implemented properly
- Batched implementations here: https://github.com/tensorflow/gan/ (Code WIP...)

HYPE

- Human eYe Perceptual Evaluation
- Amazon Mechanical Turk
- “Image exposures are in the range [100 ms, 1000 ms], derived from the perception literature”
- Exposure time can be adaptive HYPE_{time}
- Exposure time can be infinite HYPE_{∞}
- https://stanfordhci.github.io/gen-eval/

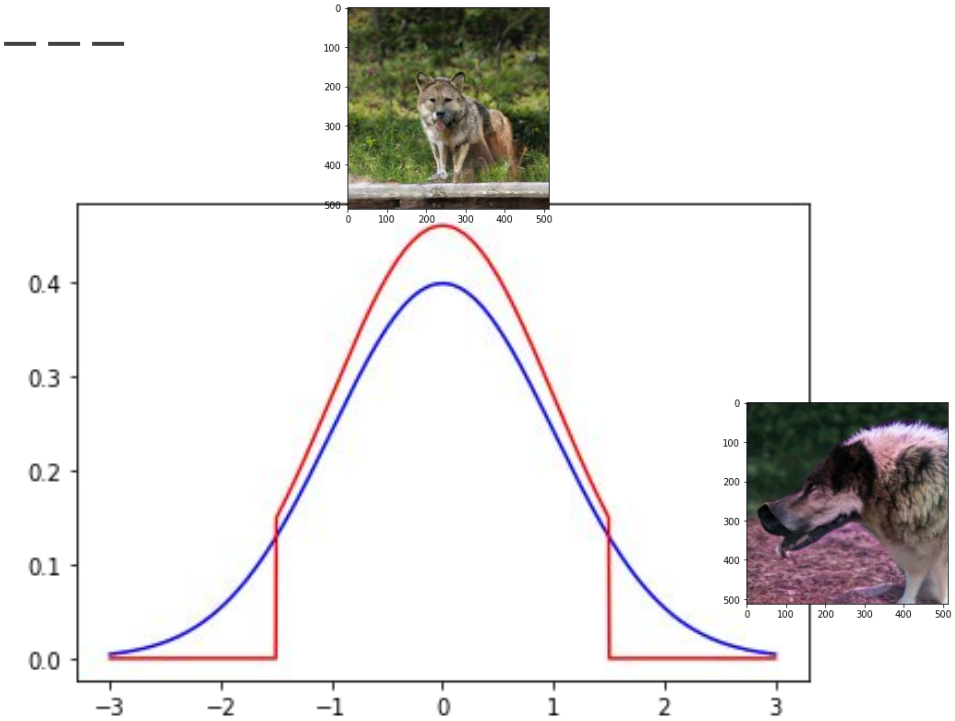
Zhou, Sharon, et al. "Hype: A benchmark for human eye perceptual evaluation of generative models." Advances in Neural Information Processing Systems. 2019.



General evaluation

- Focus on the downstream task
  - If GANs are used for augmentation for a classification task = accuracy of the classifier
  - If GANs are used to generate realistic images = HYPE
  - If GANs are used for video compression = bandwidth
  - If GANs are used for steganography = detectability using SOTA steganalysis models
  - ...

Truncation



Rank	GAN	$\mathbf{HYPE}_{\infty}$ (%)
1	$\text{StyleGAN}_{\text{trunc}}$	27.6%
2	$\text{StyleGAN}_{\text{no-trunc}}$	19.0%

Rank	GAN	$\mathbf{HYPE}_{\text{time}}$ (ms)
1	$\text{StyleGAN}_{\text{trunc}}$	363.2
2	$\text{StyleGAN}_{\text{no-trunc}}$	240.7

Zhou, Sharon, et al. "Hype: A benchmark for human eye perceptual evaluation of generative models." Advances in Neural Information Processing Systems. 2019.

- **Trade-off = diversity / fidelity**

Conclusions

Conclusions

- GANs studied
  - Vanilla GANs
  - Non Saturating GANs
  - DC-GANs
  - Conditional GANs
  - W-GANs-GP
- GANs are very promising
- Still very hacky
- Active area of research
- https://github.com/hindupuravinash/the-gan-zoo

Next video

– – –

- More on GAN Failures
- Image to image translation
  - Paired (pix2pix)
  - Unpaired (cycle GANs)
- Style GANs
- ...

End