Catherine Yeo

WEEK 1: EVALUATION OF GANs

Evaluation

- Evaluating GANs is hard b/c there's no concrete metric/goal for how realistic the output is
  - Discriminator never reaches perfection, will overfit
- 2 properties:
  1) Fidelity: quality of images (crispness, realism)
  2) Diversity: variety of images generated, want whole distr. covered

Comparing Images

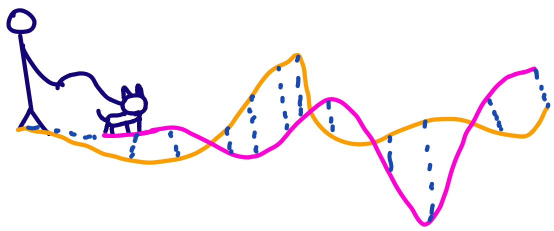
- Pixel distance : subtract pixels' hex values
  - Not reliable ⇒ if photo shifted by 1, pixel distance would be HUGE
- Feature distance : use higher-level features
  - Extract features → Compare

Extract features	Compare
2 eyes	2 eyes
1 nose	1 nose
2 legs	4 legs

  - How to extract features?
    - Pre-trained classifier ⇒ weights have encoded a lot of features
      - Take outputs of earlier layers (last pooling layer) , lop off last 2 layers
    - ImageNet
      - Inception-v3 network — super computation efficient
      - Features repr. by embeddings, compare to get feature distance

Fréchet Inception Distance (FID)

- Fréchet distance



Dog-walking analogy: what is the min. leash distance needed?

  - FD b/t normal distr.s: $d(X,Y) = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2$

- Multivariate normal FD = FID:

$$\|\mu_x - \mu_y\|^2 + \mathrm{Tr}\left(\Sigma_x + \Sigma_y - 2\sqrt{\Sigma_x \Sigma_y}\right)$$

↪ trace = sum of diagonal

  - Real (x) & fake (y) embeddings are 2 multivariate normal distr.s
  - Lower FID = closer distributions = better
  - Use large sample size to reduce noise & selection bias

Inception Score

- Keep model intact, don't lop
- Entropy: Fidelity = low entropy; Diversity = high entropy
- KL divergence

conditional distr.

$$D_{KL}\left(p(y|x) \,\|\, p(y)\right) = \boxed{p(y|x)} \log\left(\frac{\boxed{p(y|x)}}{\boxed{p(y)}}\right)$$

marginal distr.

- Inception Score (IS)

high score = good
low = bad

$$IS = \exp\left(E_{x\sim p} D_{KL}\left(p(y|x) \,\|\, p(y)\right)\right)$$

- Can be exploited by diversity ⇒ mode collapse
- Only looks at fake images
- Thus, worse than FID

Sampling & Truncation

- Truncation trick
  - Sample at test time from N distr. with tails clipped
  - Higher fidelity = sample around 0, truncate more of tails
  - Higher diversity = sample from tails, truncate less of tails
- Human evaluation still necessary for sampling

Precision & Recall

- Precision = overlap b/t fakes & reals / all fakes
  - Higher precision = better fidelity
- Recall = overlap b/t fakes & reals / all reals
  - Higher recall = better diversity

WEEK 2 : GAN DISADVANTAGES & BIAS

Analysis

PROs

- Amazing results (esp. fidelity)
- Generate quickly (fast inference)

CONs

- Lack intrinsic evaluation metrics
- Unstable training (e.g. mode collapse)
- No density estimation
- Inverting (image → noise vector) not easy

Alternatives to GANs

- Generative models : $\underset{\text{Noise}}{\xi}, \underset{\text{Class}}{Y} \rightarrow \underset{\text{Features}}{X}$ model $P(X|Y)$
- VAEs : try to minimize divergent b/t generated & real distributions
  ↳ take a real image, represent it in latent space, then use encoder & decoder

PROs

- Density estimation
- Invertible
- Stable training

CONs

- Lower quality results (blurrier)

- Autoregressive models : conditions on previous pixels to generate next pixel
- Flow models : uses invertible mappings
- Hybrid models too!

Machine Bias

- Risk assessment → COMPAS algorithm
  - Low accuracy
  - Proxies for race

Ways Bias is Introduced

- Training bias
  - Variation of who/what in data
  - Collection methods
  - Diversity of labellers
- Evaluation bias → reinforce & amplify biases from data
- Model architecture bias
- Bias can appear at <u>any</u> step
- PULSE case study
  - Upsamples low res → high res photos
  - But it upsamples POC to white faces

WEEK 3: STYLEGAN & ADVANCEMENTS

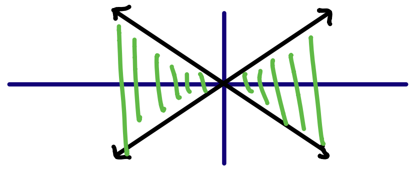
GAN Improvements

- Stability → longer training & better images

  | 4 | 3 | 7 | → high st. dev. (variation)

  | 4 | 4 | 4 | → low st. dev., mode collapse

  - Use minibatch rather than 1 number at a time
  - Enforce 1-Lipschitz continuity
    ↳ weight normalization

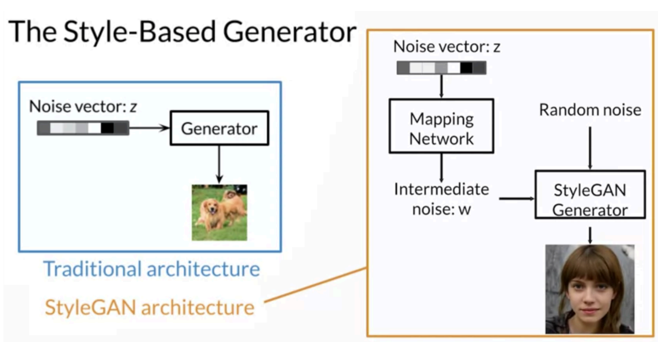


  - Moving average of weights → smoother results
  - Progressive growing
- Capacity → larger models can use higher resolution images
- Diversity → more generated variety
  - Minibatch helps too

StyleGAN Overview

GOALS { greater fidelity
increased diversity of outputs
more control over image features

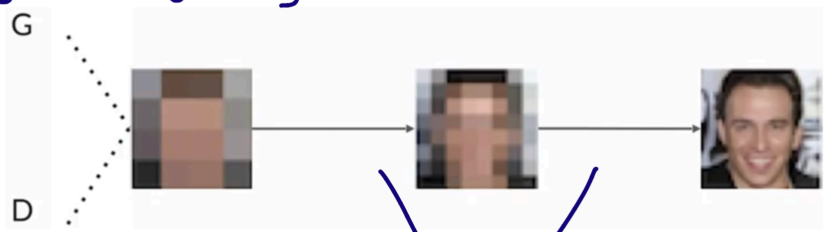
- Styles: coarse, middle, fine variations in image
  ↳ face shape ↳ hair color
- Generator:



random Gaussian noise

AdaIN (adaptive instance norm.)

- Progressive growing



doubling (but slow)

Progressive Growing

- Ex: (G) start with 4 x 4 image

  double 8 x 8 , extra convolutional layer → higher res

  upsample 2x + learned parameters

  ⋮ scheduled intervals

  1024 x 1024

- (G) Upsample 2x $1-\alpha$

  Double layer α

- (D) Downsample 0.5x $1-\alpha$

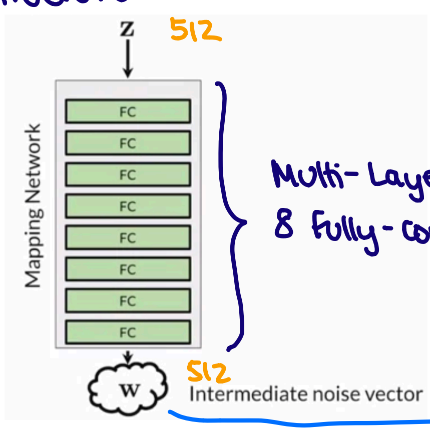
  . . . α

- TL; DR { gradually 2x resolution
  faster + more stable

Noise Mapping Network

- Structure

512



512

Multi-Layered Perceptron w/ 8 Fully-connected layers

4 x 4
$\boxed{8 \times 8}$
16 x 16
⋮
1024 x 1024

Upsample
Convolution
AdaIN
Convolution
AdaIN

- Noise is from z-space entanglement (no 1-to-1 mappings)
- Intermediate w can learn 1-to-1 mappings, less entangled noise space
  ↳ becomes inputs to generator

Adaptive Instance Normalization (AdaIN)

- Instance Normalization : looks at 1 example at a time, use its μ & σ
- AdaIN

  1) $\dfrac{x_i - \mu(x_i)}{\sigma(x_i)}$ $(\mu = 0,\ \sigma = 1)$ — Normalize examples

  2) Apply adaptive styles using w (shifting values) — Transfer style info from w to image

$$\boxed{\mathrm{AdaIN}(x_i, y) = y_{s,i} \frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i}}$$

Style Mixing & Stochastic Noise

- Sample $z_1 \rightarrow w_1 \rightarrow \cdots \rightarrow$ AdaIN layers
  ↳ coarse
  $z_2 \rightarrow w_2 \rightarrow \ . \ . \ .$
  ↳ fine

- Inject noise into later layers → finer details
  earlier → coarse