Conditional GANs, Image-to-Image Translation

Jun-Yan Zhu
16-726, Spring 2022
Ideal models (Dream)
   Pros: good sample, fast sample, Exact/fast likelihoods
   good coverage, easy to training, learn low-dimensional latent representation.

Autoregressive models
   Pros: Exact likelihoods, good coverage
   Cons: Slow to evaluate or sample

VAEs
   Pros: fast to sample, fast to train, good coverage
   Cons: Blurry samples (in practice)

GANs
   Pros: fast to sample, fast to train, good samples
   Cons: No likelihoods (density), bad coverage (mode collapse)

Flow-based models
   Pros: fast to sample, exact likelihoods
   Cons: memory-intensive; slow training; limited choices for generators,
   high-dimensional codes

Diffusion models
   Pros: good samples, good coverage
   Cons: slow training, slow sampling
Which model is better?

• It depends on your applications
  • Synthesis
  • Classification
  • Density estimation

• Which model is easier to train?

• Which model is faster (training & inference)?
Conditional GANs, Image-to-Image Translation

Jun-Yan Zhu
16-726, Spring 2021

© https://affinelayer.com/pixsrv/, pix2pix [Isola et al., 2016]
Problem Statement

**Goal:** synthesize a photograph given an input image

Input:
- Sky
- Mountain
- Sea
- Rock

Output:

---

Sky

Sea

Mountain

Rock

---
Problem Statement

Goal: synthesize a photograph given an input image
Early work (Example-based)

Semantic Photo Synthesis [Johnson et al., Eurographics 2006]

Sketch2Photo [Tao et al., SIGGRAPH Asia 2009]
Semantic Photo Synthesis

Semantic Photo Synthesis [EG’06]

Semantic Photo Synthesis
Semantic Photo Synthesis

Learning-based methods
Loss functions for Image Synthesis

What is a good objective $\mathcal{L}$?
- What is a good loss?
- How to calculate it efficiently?
- How to collect data $(x, y)$?

Problem Statement

$$\operatorname{arg min}_{G} \mathcal{L}(G(x), y)$$

Generator $G$  
Learnable rendering  
Output Image $G(x)$
Designing Loss Functions

$\text{L2 regression } \arg \min_G \mathbb{E}[||G(x) - y||]$
Designing Loss Functions

Image colorization

Super-resolution

L2 regression

L2 regression
Designing Loss Functions

Image colorization

Super-resolution

Classification Loss:
Cross entropy objective, with colorfulness term

Feature/Perceptual loss
Deep feature matching objective

[Zhang et al. 2016], [Gatys et al., 2016], [Johnson et al. 2016], [Dosovitskiy and Brox. 2016]
“Perceptual Loss”


CNNs as a Perceptual Metric

\[ F \]

Normalize, Subtract

\[ L_2 \text{norm, Spatial average} \]

\( G(x) \)  \( y \)  \( d_0 \)

**CNNs as a Perceptual Metric**

F is a deep network (e.g., ImageNet classifier)

**Perceptual Loss**

\[
\arg\min_G \mathbb{E}_{(x,y)} \sum_{i=1}^{N} \lambda_i \frac{1}{M_i} \left\| F^{(i)}(G(x)) - F^{(i)}(y) \right\|^2_2
\]

The number of elements in the (i)-th layer
Learning with Perceptual Loss

Training objective: \[
\arg \min_G \mathbb{E}_{(x,y)} \sum_{i=1}^{N} \lambda_i \frac{1}{M_i} \left\| F^{(i)}(G(x)) - F^{(i)}(y) \right\|^2
\]

CRN [Chen and Koltun, 2017]
Generated images

Universal loss?
Generative Adversarial Network (GANs)

Real vs. Fake

Classifier

Real photos

Generated images

[Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, Bengio 2014]
Goodfellow et al. 2014
A two-player game:
- G tries to generate fake images that can fool D.
- D tries to detect fake images.

[Goodfellow et al. 2014]
Learning objective

\[
\min_G \max_D \mathbb{E}_x \left[ \log(1 - D(G(x))) \right] + \mathbb{E}_y \left[ \log D(y) \right]
\]

[Goodfellow et al. 2014]
Learning objective

$$\min_G \max_D \mathbb{E}_x \left[ \log(1 - D(G(x))) \right] + \mathbb{E}_y \left[ \log D(y) \right]$$

[Goodfellow et al. 2014]
Learning objective

$$\min_G \max_D \mathbb{E}_x \left[ \log (1 - D(G(x))) \right] + \mathbb{E}_y \left[ \log D(y) \right]$$

[Goodfellow et al. 2014]
Learning objective

\[
\min_G \max_D \mathbb{E}_x [\log(1 - D(G(x)))] + \mathbb{E}_y [\log D(y)]
\]

Pix2pix [Isola et al., 2016]
Learning objective

\[
\min_G \max_D \mathbb{E}_x[\log(1 - D(G(x)))] + \mathbb{E}_y[\log D(y)]
\]

Pix2pix [Isola et al., 2016]
Learning objective

$$\min_G \max_D \mathbb{E}_x \left[ \log(1 - D(x, G(x))) \right] + \mathbb{E}_{x,y} \left[ \log D(x, y) \right]$$

Pix2pix [Isola et al., 2016]
pix2pix Generator (U-Net)

U-Net [Ronneberger et al.]: popular CNN backbone for biomedical image segmentation

U-Net: preserve high-frequency information (e.g., edge) of the input image.

Encoder-decoder: lose high-frequency details due to the information bottleneck
pix2pix Generator (U-Net)

Generator design is critical for image quality. cGAN (conditional GANs) loss: capture realism. L1 loss stabilizes training (faster convergence)
pix2pix Discriminator (PatchGAN)

- Rather than penalizing if output image looks fake, penalize if each overlapping patches looks fake.
- Focus on local visual cues (color, textures).
- Global structure: the input image has already encoded global structure. L1 loss can help as well.

Advantages
- Faster, fewer parameters
- More supervised observations
- Applies to arbitrarily large images
# edges2cats

[Christopher Hesse]

Ivy Tasi @ivymyt

@gods_tail

Vitaly Vidmirov @vvid

https://affinelayer.com/pixsrv/
Input: Sketch → Output: Photo
Input: Grayscale → Output: Color

Real or fake pair?
Automatic Colorization with pix2pix

Data from [Russakovsky et al. 2015]
Automatic Colorization with pix2pix

Data from [Russakovsky et al. 2015]
Input: **Text**  →  Output: **Photo**

Text-to-Image Synthesis

StackGAN, StackGAN++ [Zhang et al., 2016 and 2017], AttnGAN [Xu et al., 2018]
Input: Class → Output: Photo

Class-conditional GANs

cGANs [Mirza and Osindero. 2014], SAGAN [Zhang et al., 2018], BigGAN [Brock et al., 2019]

StyleGAN-XL [Sauer et al., 2022]
Class-conditional Discriminator

(a) cGANs, input concat (Mirza & Osindero, 2014)
Adversarial loss

(b) cGANs, hidden concat (Reed et al., 2016)
Adversarial loss

(c) AC-GANs (Odena et al., 2017)
Adversarial classification loss

(d) (ours) Projection

\[ f(x, y) := y^T V \phi(x) + \psi(\phi(x)) \]

x: image, y: class labels (one-hot vector), \( \phi, \phi', \psi \): neural networks

Learnable matrix Projection Discriminator [Miyato and Koyama, ICLR 2018]
Class-conditional Discriminator

Projection Discriminator [Miyato and Koyama, ICLR 2018]
BigGAN

(a) 128×128  (b) 256×256  (c) 512×512  (d)
Real or fake pair?

Learning objective:

$$\min_G \max_D \mathbb{E}_x [\log(1 - D(x, G(x))] + \mathbb{E}_{x, y} [\log D(x, y)]$$

Pix2pix [Isola et al., 2016]
Limitations

• One-to-one mapping.
• Low-resolution output.
• Requires paired training data
Improving Conditional GANs

- Multimodal synthesis.
- High-resolution synthesis.
- Model training without pairs (next lecture)
Synthesizing Multiple Results

Multi-agent Diverse GANs [Ghosh et al., CVPR 2018]
Synthesizing Multiple Results

Night input  Day output 1  Day output 2  Day output 3

Multi-agent Diverse GANs [Ghosh et al., CVPR 2018]
Synthesizing Multiple Results

Multi-agent Diverse GANs [Ghosh et al., CVPR 2018]

Discriminator: K+1 classification
Generator: fool D to classify fake as real

real OR from $G_1 \ G_2 \ \cdots \ \ G_K$
Synthesizing Multiple Results

Night input  
Day output 1  
Day output 2  
Day output 3

Multi-agent Diverse GANs [Ghosh et al., CVPR 2018]
Synthesizing Multiple Results

\[ G(x, E(x)) \sim x \]
\[ E(G(x, z)) \sim z \quad \text{KL Divergence} \]

VAE-GAN [Larsen et al., 2016], BicycleGAN [Zhu et al., 2017]
Synthesizing Multiple Results

BicycleGAN [Zhu et al., 2017]
Synthesizing Multiple Results

BicycleGAN [Zhu et al., 2017]
Synthesizing Multiple Results

\[ G(x, z_1) \]

\[ \max_G L_z(G) = \mathbb{E}_{z_1, z_2} \left[ \min \left( \frac{\|G(x, z_1) - G(x, z_2)\|}{\|z_1 - z_2\|}, \tau \right) \right], \]

DSGAN [Zhu et al., 2017]
Synthesizing Multiple Results

\[ G(x, z_1) \quad \text{and} \quad G(x, z_2) \]

\[
\max_G \mathcal{L}_z(G) = \mathbb{E}_{z_1, z_2} \left[ \min \left( \frac{\|z_1 - z_2\|}{\tau}, 1 \right) \right]
\]

DSGAN [Yang et al., 2019]
Synthesizing Multiple Results

\[
\max_G \mathcal{L}_z(G) = \mathbb{E}_{z_1, z_2} \left[ \min \left( \frac{\|G(x, z_1) - G(x, z_2)\|}{\|z_1 - z_2\|}, \tau \right) \right]
\]

DSGAN [Yang et al., 2019]
Improving Conditional GANs

- Multimodal synthesis.
- High-resolution synthesis.
- Model training without pairs (next lecture)
The Curse of Dimensionality

Pix2pix output
pix2pixHD

Objective: Multi-scale GANs loss + Perceptual Loss + Feature Matching Loss (with Discriminator’s features)

Image Pyramid [Burt and Adelson, 1987]
Also see [Zhang et al., 2017]
[Karras et al., 2018]
pix2pixHD: 2048×1024
Conditional Image Synthesis in the Wild

pix2pixHD [Wang et al., 2018]
pix2pixHD [Wang et al., 2018]

input

sky

output

grass


Problem with standard networks

rock, water, moss, …
Problem with standard networks

grass

conv

normalization
normalization

normalization

normalization
SPADE (SPAtially ADaptive DEnormalization)
SPADE (SPAtially ADaptive DEnormalization)

Batch Norm (Ioffe et al. 2015)

\[ y = \frac{x - \mu}{\sigma} \cdot \gamma + \beta \]

See other adaptive/conditional normalization: conditional BN (Dumoulin et al.), AdaIN (Huang and Belongie), SFT (Wang et al.)
Semantic Control
Semantic Control
Semantic Control
Style Control
Style Control

Style Manipulation
Style Control

Style Manipulation
SIGGRAPH 2019 Real-time Live! “Best of Show Award" and "Audience Choice Award"
By Darek Zabrocki, Concept Designer and Illustrator
Thank You!

16-726, Spring 2022