Generative Adversarial Networks
For Image to Image Translation

Sagie Benaim
Tel Aviv University
Generative Modeling: Density Estimation

Training Data

Density Function
Generative Modeling: Sample Generation

Training Data (CelebA)  Sample Generator (Karras et al, 2017)
Adversarial Nets Framework

$D(x)$ tries to be near 1

Differentiable function $D$

$x$ sampled from data

$D$ tries to make $D(G(z))$ near 0,
$G$ tries to make $D(G(z))$ near 1

$x$ sampled from model

Differentiable function $G$

Input noise $z$

(Goodfellow et al., 2014)
Conditional GAN
Image to Image Translation

- **Monet** ↔ **Photos**
- **Zebras** ↔ **Horses**
- **Summer** ↔ **Winter**

- **Monet** → **photo**
- **zebra** → **horse**
- **summer** → **winter**

- **photo** → **Monet**
- **horse** → **zebra**
- **winter** → **summer**

- **Photograph**
- **Monet**
- **Van Gogh**
- **Cezanne**
- **Ukiyo-e**
<table>
<thead>
<tr>
<th></th>
<th>Supervised</th>
<th>Unsupervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unimodal</td>
<td>Pix2pix, CRN, SRGAN</td>
<td>DistanceGAN, CycleGAN, DiscoGAN, DualGAN, UNIT, DTN, StarGAN, OST</td>
</tr>
<tr>
<td>Multimodal</td>
<td>pix2pixHD, BicycleGAN</td>
<td>MUNIT, Augmented CycleGAN</td>
</tr>
</tbody>
</table>
Fully Supervised: pix2pix

Conditional GAN

\[ G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G). \]

[Isola et al., CVPR 2017]
Isola et al., CVPR 2017
Unsupervised: Circular GANs

**DiscoGAN**: “Learning to Discover Cross-Domain Relations with Generative Adversarial Networks”. Kim et al. ICML’17.


Cycle-Consistent Adversarial Networks

[X]

[Mark Twain, 1903]

[Y]

[D_Y]

[Zhu et al., ICCV 2017]
Cycle Consistency Loss

\[
\|F(G(x)) - x\|_1 \quad \text{Reconstruction error}
\]

\[
\|G(F(y)) - y\|_1 \quad \text{Reconstruction error}
\]

See similar formulations [Yi et al. 2017], [Kim et al. 2017]

[Zhu et al., ICCV 2017]
Collection Style Transfer

Photograph @ Alexei Efros

Monet

Van Gogh

Cezanne

Ukiyo-e
DistanceGAN

• A pair of images of a given distance are mapped to a pair of outputs with a similar distance
• $|x_i - x_j|_1$ and $|G(x_i) - G(x_j)|_1$ are highly correlated.

![Diagram](image)

$|x_1 - x_2|_1 \sim |G(x_1) - G(x_2)|_1$

Benaim et al., NIPS 2017
Motivating distance correlations

Analysis of CycleGAN’s horse to zebra results

Benaim et al., NIPS 2017
Less Supervision: Only a single image in domain A

Many unmatched samples in domain B

+ One sample $x$ in domain A

→ Analogue of $x$ in B

One Shot Unsupervised Cross Domain Translation (NeurIPS 2018)
Modeling multiple possible outputs

Input

Possible outputs
**BiCycleGAN** [Zhu et al., NIPS 2017] (c.f. InfoGAN [Chen et al. 2016])

**MAD-GAN** [Ghosh et al., CVPR 2018]
MUNIT: Multimodal Translation

(a) Auto-encoding

(b) Translation

Huang et al., ECCV 2018
Sketch to Image Translation

(a) edges $\leftrightarrow$ shoes

(b) edges $\leftrightarrow$ handbags
Animal Image Translation

(a) house cats $\rightarrow$ big cats
(b) big cats $\rightarrow$ house cats
(c) house cats $\rightarrow$ dogs
(d) dogs $\rightarrow$ house cats
(e) big cats $\rightarrow$ dogs
(f) dogs $\rightarrow$ big cats

Huang et al., ECCV 2018
"Emerging Disentanglement in Auto-Encoder Based Unsupervised Image Content Transfer", ICLR 2019
Domain B

Separate Encoder

Decoder

Domain B

Common Encoder

Adversarial Loss

Domain A

Separate Encoder

0

Decoder

Domain A

Zero Loss

Reconstruction Loss

Reconstruction Loss
Adversarial Loss

Domain B

Common Encoder

Discriminator

Domain A

Is encoding from domain A or B?
Other Domains?

- Audio Separation: Training data consists of a set of samples of mixed music and an unmatched set of instrumental music.
- Given a mixed sample, wish to separate the voice from the background instrumental music.
- After mapping the audio sample to a Spectrogram, can subtract the “background” from the “mixed” sample in “pixel space”, to get the “voice” only sample.
- Samples at: https://sagiebenaim.github.io/Singing/

"Semi-Supervised Monaural Singing Voice Separation With a Masking Network Trained on Synthetic Mixtures." ICASSP 2019
Video to Video

• Use GAN to generate each frame in a video
• Use optical flow to further constrain the generator
• Samples at: https://github.com/NVIDIA/vid2vid

"High Resolution photorealistic video to video translation." NeurIPS 2018
Many More Applications

• Many other Vision Applications: Photo Enhancement, Image Dehazing
• Medical Imaging and Biology [Wolterink et al., 2017]
• Voice conversion [Fang et al., 2018, Kaneko et al., 2017]
• Cryptography [CipherGAN: Gomez et al., ICLR 2018]
• Robotics
• NLP: Unsupervised machine translation.
• NLP: Text style transfer.
• ...
Thank You! Questions?