Generative Adversarial Networks For Image to Image Translation

Sagie Benaim
Tel Aviv University
Generative Modeling: Density Estimation

Training Data

Density Function
Adversarial Nets Framework

\[ D(x) \text{ tries to be near 1} \]
\[ D \text{ tries to make } D(G(z)) \text{ near 0, } \]
\[ G \text{ tries to make } D(G(z)) \text{ near 1} \]

\[ D \]
\[ x \text{ sampled from data} \]
\[ x \text{ sampled from model} \]
\[ \text{Differentiable function } D \]
\[ \text{Differentiable function } G \]

\[ \text{Input noise } z \]

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

Monet ➔ Photos ➔ Monet

Zebras ➔ Horses ➔ zebra ➔ horse

Summer ➔ Winter ➔ summer ➔ winter

photo ➔ Monet ➔ horse ➔ zebra ➔ winter ➔ summer

Photograph ➔ Monet ➔ Van Gogh ➔ Cezanne ➔ Ukiyo-e
<table>
<thead>
<tr>
<th>Modality</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>
Paired

\[ x_i \]

\[ y_i \]

Unpaired

\[ X \]

\[ Y \]
Fully Supervised: pix2pix

Conditional GAN

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

[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

[Mark Twain, 1903]

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

\[
\begin{align*}
\forall x &\quad G(x) &\quad F(G(x)) &\quad y &\quad F(y) &\quad G(F(x)) \\
&\xrightarrow{G} \quad \hat{Y} &\quad \hat{x} &\quad \hat{X} &\quad \hat{y} \\
&\downarrow D_Y(G(x)) &\quad D_G(F(x))
\end{align*}
\]

Reconstruction error

\[
\|F(G(x)) - x\|_1, \quad \|G(F(y)) - y\|_1
\]

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.

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 $\rightarrow$ 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 ↔ shoes

(b) edges ↔ handbags

Huang et al., ECCV 2018
Animal Image Translation

(a) house cats → big cats
(b) big cats → house cats
(c) house cats → dogs
(d) dogs → house cats
(e) big cats → dogs
(f) dogs → big cats

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

Separate Encoder

Common Encoder

Adversarial Loss

Domain A

Separate Encoder

Decoder

Reconstruction Loss

Zero Loss

Adversarial Loss

Domain B

Decoder

Reconstruction Loss

Domain A

Decoder
Adversarial Loss

Domain B

Common Encoder

Discriminator

Is encoding from domain A or B?

Domain A
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 the 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/](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 from 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?