Image-to-Image Translation

Ming-Yu Liu
NVIDIA

Ting-Chun Wang
NVIDIA
A Short Introduction to GANs

Ting-Chun Wang
NVIDIA
Recent Advances in Image Generation

Unconditional Image Synthesis

- Prog. GAN [Karras et al.]
- BigGAN [Brock et al.]
- StyleGAN [Karras et al.]

Image-to-Image Synthesis

- pix2pix [Isola et al.]
- pix2pixHD [Wang et al.]
- GauGAN [Park et al.]

Text-to-Image Synthesis

- TextGAN [Reed et al.]
- StackGAN [Zhang et al.]
- AttnGAN [Xu et al.]

Video Synthesis

- TGAN [Saito et al.]
- MoCoGAN [Tulyakov et al.]
- vid2vid [Wang et al.]
Recent Advances in Image Generation

- **Unconditional Image Synthesis**
  - BigGAN [Brock et al.]
  - StyleGAN [Karras et al.]

- **Image-to-Image Synthesis**
  - pix2pix [Isola et al.]
  - pix2pixHD [Wang et al.]
  - GauGAN [Park et al.]

- **Text-to-Image Synthesis**
  - TextGAN [Reed et al.]
  - StackGAN [Zhang et al.]
  - AttnGAN [Xu et al.]

- **Video Synthesis**
  - TGAN [Saito et al.]
  - MoCoGAN [Tulyakov et al.]
  - vid2vid [Wang et al.]
Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two different image domains

• e.g. day-time image domain & night-time image domain

Let $x_1 \in \mathcal{X}_1$

$\text{I2I}$: the problem of translating $x_1$ to a corresponding image $x_2 \in \mathcal{X}_2$

• Correspondence can mean different things in different contexts
Examples and Use Cases

- Low-res to high-res
- Blurry to sharp
- Thermal to color
- Synthetic to real

- LDR to HDR
- Noisy to clean
- Image to painting

- Day to night
- Summer to winter

- Bad weather to good weather
- Greyscale to color
- …
Prior Works

• Image translation has been studied for decades
• Different approaches have been exploited, including
  • Filtering-based
  • Optimization-based
  • Dictionary learning-based
  • Deep learning-based
  • GAN-based
Prior Works

• Image translation has been studied for decades
• Different approaches have been exploited, including
  • Filtering-based
  • Optimization-based
  • Dictionary learning-based
  • Deep learning-based
  • GAN-based
Why are GANs useful for I2I?

- Hand-crafted features → Deep Networks
- Hand-crafted objective function → Generative Adversarial Networks
Generative Adversarial Networks (GANs)

- Forget about designing a perceptual loss
- Let’s train a new network to differentiate real and fake images
GAN Objective

Solving a minimax problem

$$\min_G \max_D \quad E_{x \sim p_X} [\log D(x)] + E_{z \sim p_Z} [\log (1 - D(G(z)))]$$
GAN Objective

Solving a minimax problem

For discriminator $D$:

$$\min_G \max_D \mathbb{E}_{x \sim p_X} \left[ \log D(x) \right] + \mathbb{E}_{z \sim p_Z} \left[ \log(1 - D(G(z))) \right]$$

- real samples
- generated samples
GAN Objective

Solving a minimax problem

For Generator G:

\[
\min_G \max_D E_{x \sim p_x} [\log D(x)] + E_{z \sim p_z} [\log (1 - D(G(z)))]
\]
GAN Objective

Solving a minimax problem

$$\min_G \max_D E_{x \sim p_X} [\log D(x)] + E_{z \sim p_Z} [\log (1 - D(G(z)))]$$

Training: done by alternating two stochastic gradient updates

Update G: $$\max_G E_{z \sim p_N} [\log D(G(z))]$$

Update D: $$\max_D E_{x \sim p_X} [\log D(X)] + E_{z \sim p_N} [\log (1 - D(G(z)))]$$
Unconditional vs. Conditional GANs

Unconditional: $z \sim \mathcal{N} \rightarrow G \rightarrow G(z) \rightarrow D
\quad x \in \mathcal{X} \rightarrow D

Conditional: $x_1 \sim \mathcal{X}_1 \rightarrow F \rightarrow F(x_1) \rightarrow D
\quad x_2 \sim \mathcal{X}_2 \rightarrow D
Conditional GAN for Image Translation

• Conditional GAN loss alone is insufficient for image translation
  • No guarantee the translated image is related to the source image
  • Generator can just completely ignore source images

• This can be easily fixed in the supervised setting
  • Where ground truth image pairs before/after translation are available

\[
\text{Dataset} = \{ (x_1^{(1)}, x_2^{(1)}), (x_1^{(2)}, x_2^{(2)}), \ldots, (x_1^{(N)}, x_2^{(N)}) \}\]
Supervised Image-to-Image Translation

\[ x_1 \sim \mathcal{X}_1 \xrightarrow{F} x = F(x_1) \rightarrow \text{fake} \]

\[ x_2 \rightarrow \text{real} \]

• Supervisedly relating \( x = F(x_1^{(i)}) \) to \( x_2^{(i)} \)

• Ledig et al (CVPR’17): Adding content loss

\[ \|x - x_2^{(i)}\|_2 + \|\text{VGG}(x) - \text{VGG}(x_2^{(i)})\|_2 \]
SRGAN


bicubic (21.59dB/0.6423)  SRResNet (23.53dB/0.7832)  SRGAN (21.15dB/0.6868)  original
Supervised Image-to-image Translation

\[ x_1 \sim \mathcal{X}_1 \xrightarrow{F} x = F(x_1) \rightarrow \text{fake} \]
\[ x_2 \rightarrow \text{real} \]

- Supervisedly relating \( x = F(x_1^{(i)}) \) to \( x_2^{(i)} \)
- Isola et al (CVPR’17): Learning a joint distribution
  Discriminator sees both input and output images
  \[
  \max_F E_{pX_1} [\log(D(x_1, F(x_1)))]
  \]
Pix2Pix

Unsupervised Image-to-image Translation

• Corresponding images could be expensive to get
• In the unsupervised setting
  • No correspondence between the two datasets

\[
\text{Dataset}_1 = \{x_1^{(n_1)}, x_1^{(n_2)}, \ldots, x_1^{(n_N)}\}
\]

\[
\text{Dataset}_2 = \{x_2^{(m_1)}, x_2^{(m_2)}, \ldots, x_2^{(m_M)}\}
\]

• Need additional constraints/assumptions for learning the translation
SimGAN

\[ x_1 \sim \mathcal{X}_1 \rightarrow F \rightarrow x = F(x_1) \rightarrow D \]

- Srivastava et al. (CVPR’17): adding cross-domain content loss

\[
\max_F E_{p_{x_1}} \left[ \log D(F(x_1)) - \lambda \| F(x_1) - x_1 \|_1 \right]
\]
Cycle Constraint

\[ x_1 \sim \mathcal{X}_1 \xrightarrow{F_{12}} \hat{x}_2 = F_{12}(x_1) \xrightarrow{F_{21}} F_{21}(F_{12}(x_1)) \]

- Learning a two-way translation
- DiscoGAN by Kim et al. (ICML’17)
- CycleGAN by Zhu et al. (ICCV’17)

\[
\max_{F_{12}, F_{21}} \mathcal{E}_{p x_1} \left[ \log(D_2(F_{12}(x_1))) - \lambda \|F_{21}(F_{12}(x_1)) - x_1\|_p^p \right] + \\
\mathcal{E}_{p x_2} \left[ \log(D_1(F_{21}(x_2))) - \lambda \|F_{12}(F_{21}(x_2)) - x_2\|_p^p \right]
\]
CycleGAN: Unsupervised Image-to-Image Translation
Image Translation Results


Results on Unsupervised RGB-Image-to-Thermal-Image Translation. Left: input color image. Right: Output thermal image.

Results on Unsupervised Day-Image-to-Night-Image Translation. Left: input day image. Right: Output night image.

Results on Unsupervised Night-Image-to-Day-Image Translation. Left: input night image. Right: Output day image.
Image Translation Results

Results on Unsupervised Sunny-Image-to-Rainy-Image Translation. Left: input sunny image. Right: Output rainy image.

Results on Unsupervised Rainy-Image-to-Sunny-Image Translation. Left: input rainy image. Right: Output sunny image.

Back View  Front View  Left View  Right View

Foggy image to clear sky image
Attribute-based Face Image Translation
Image Translation Results

Input +Blondhair

Input +Eyeglasses

Input -Goatee

Input -Smiling
Image Translation Results

Input +Blondhair

Input +Eyeglasses

Input +Goatee

Input +Smiling
## Improving GAN Training

<table>
<thead>
<tr>
<th>Tricks</th>
<th>New objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Label smoothing</td>
<td>• EBGAN</td>
</tr>
<tr>
<td>• Historical batches</td>
<td>• LSGAN</td>
</tr>
<tr>
<td>• ...</td>
<td>• WGAN</td>
</tr>
<tr>
<td>• ...</td>
<td>• BEGAN</td>
</tr>
<tr>
<td>• ...</td>
<td>• ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Surrogate or auxiliary objective</th>
<th>Network architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>• UnrolledGAN</td>
<td>• LAPGAN</td>
</tr>
<tr>
<td>• WGAN-GP</td>
<td>• Stacked GAN</td>
</tr>
<tr>
<td>• DRAGAN</td>
<td>• ...</td>
</tr>
<tr>
<td>• ...</td>
<td>• ...</td>
</tr>
</tbody>
</table>
Wasserstein GAN


Replace classifier with a critic function

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>GAN</th>
<th>WGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\max_D E_{x \sim p_X}[\log D(x)] + E_{z \sim p_Z}[\log(1 - D(G(z)))]$</td>
<td>$\max_D E_{x \sim p_X}[D(x)] - E_{z \sim p_Z}[D(G(z))]$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Generator</th>
<th>GAN</th>
<th>WGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\max_G E_{z \sim p_Z}[\log D(G(z))]$</td>
<td>$\max_G E_{z \sim p_Z}[D(G(z))]$</td>
</tr>
</tbody>
</table>
Wasserstein GAN


GAN: minimize Jensen-Shannon divergence between $p_X$ and $p_{G(Z)}$

$$JS(p_X \| p_{G(Z)}) = KL(p_X \| \frac{p_X + p_{G(Z)}}{2}) + KL(p_{G(Z)} \| \frac{p_X + p_{G(Z)}}{2})$$

WGAN: minimize earth mover distance between $p_X$ and $p_{G(Z)}$

$$EM(p_X, p_{G(Z)}) = \inf_{\gamma \in \Pi(p_X, p_{G(Z)})} E_{(x,y) \sim \gamma}[\|x - y\|]$$
GAN vs. WGAN

- In this example
  - GAN
    - uniform (JS) distance across all space
    - gradient = 0
  - WGAN
    - smaller (EM) distance when closer to GT
    - has gradient toward GT
Disadvantage of WGAN

- Needs to ensure discriminator is 1-Lipschitz

\[ \|D(x) - D(y)\| \leq K \|x - y\| \]

- i.e., gradient is bounded everywhere and doesn’t explode
- Realized by weight clipping
WGAN-GP


Instead of weight clipping, apply **gradient penalty**

\[
\min_G \max_D E_{x \sim p_X} [D(x)] - E_{z \sim p_Z} [D(G(Z))] + \lambda E_{y \sim p_Y} [(\|\nabla_y D(y)\|_2 - 1)^2]
\]

\[
y = ux + (1 - u)G(z)
\]

\(y\): imaginary samples (between real and fake)

Optimal critic has unit gradient norm almost everywhere

<table>
<thead>
<tr>
<th>DCGAN</th>
<th>LSGAN</th>
<th>WGAN (clipping)</th>
<th>WGAN-GP (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline ((G): DCGAN, (D): DCGAN)</td>
<td>![DCGAN images]</td>
<td>![LSGAN images]</td>
<td>![WGAN (clipping) images]</td>
</tr>
</tbody>
</table>
Spectral Normalization


• Lipschitz constant of a linear function is its largest singular value (spectral norm)
  \[\|Ax\| \leq K\|x\|\]

• Spectral normalization: Replaces every weight \(W\) with \(W / \sigma(W)\)
  • \(\sigma\): largest singular value of \(W\)
  • Ensures discriminator gradient is always bounded

• Computing \(\sigma\) during training
  • Direct computation is very time consuming
  • Fast approximation using power iteration
Evaluation Metrics

• Inception Score (IS)
  • Each generated image should have a distinct label
  • Overall set of generated images should have diverse labels
  • The larger the distance between these two, the better

• Fréchet Inception Distance (FID)
  • Use inception network to extract features from images
  • Model real/fake features with two multivariate Gaussians
  • The lower the distance between these two, the better

• Human Evaluation