One-Shot Unsupervised Cross Domain Translation

Sagie Benaim and Lior Wolf
NeurIPS 2018
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

Van Gogh ➔ Cezanne

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>
Paired

\[ x_i, y_i, \ldots \]

Unpaired

\[ X, Y, \ldots \]
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]
Partially Supervised Alignment

- “Unsupervised Cross-Domain Image Generation” Taigman et al.
Unsupervised Alignment

- Highly related domains
  - “Unsupervised Image-to-Image Translation Networks” Liu et al.
Circular GANs

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


Circular GANs

Circular GANs (DiscoGAN, CycleGAN, DualGAN)

\[ x \sim F(G(x)) \]
\[ y \sim G(F(y)) \]
Generative Modeling:
Sample Generation

Training Data
(CelebA)

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

\[ D(x) \text{ tries to be near 1} \]

Differentiable function \( D \)

\[ x \text{ sampled from data} \]

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

\[ D \]

\[ x \text{ sampled from model} \]

Differentiable function \( G \)

\[ \text{Input noise } z \]

(Goodfellow et al., 2014)
Building Block: Conditional GAN

\[ \mathcal{L}_{GAN}(G_{AB}, D_B, \hat{p}_A, \hat{p}_B) = \mathbb{E}_{x_B \sim \hat{p}_B} [\log D_B(x_B)] + \mathbb{E}_{x_A \sim \hat{p}_A} [\log(1 - D_B(G_{AB}(x_A))] \]

- Other GAN variants can be used: w-gan, improved w-gan, BEGAN, etc.
Our Contribution: Only a single image in domain A

Many unmatched samples in domain B

+ One sample x in domain A

→ Analogue of x in B
Phase I
Phase I

\[ \mathcal{L}_{REC_B} = \sum_{s \in P(\Lambda)} \| G_B(E_B(s)) - s \|_1 \]

\[ \mathcal{L}_{VAE_B} = \sum_{s \in P(\Lambda)} KL(E_B \circ P(\Lambda) \| \mathcal{N}(0, I)) \]
Phase I

\[
\mathcal{L}_{GAN_B} = \sum_{s \in P(\Lambda)} -\ell(D_B(G_B(E_B(s))), 0)
\]

\[
\mathcal{L}_{DB} = \sum_{s \in P(\Lambda)} +\ell(D_B(G_B(E_B(s))), 0) + \ell(D_B(s), 1)
\]
• Shared Latent Space assumption (UNIT Liu et al, CoGAN Liu et al, etc): Upper layers of the encoder and lower layers of the decoder should be shared to achieve successful translation.
Phase I

- Shared Latent Space assumption (UNIT Liu et al, CoGAN Liu et al, etc): Upper layers of the encoder and lower layers of the decoder should be shared to achieve successful translation.

- In fact, as we only have a single sample in A, these layers, represented by the shared encoder (Es) and shared decoder (Gs) can be trained with domain B samples only
Phase II

1. $\mathcal{L}_{REC_A} = \sum_{s \in P(x)} \| T_{AA}(s) - s \|_1$
Phase II

1. $\mathcal{L}_{REC_A} = \sum_{s \in P(x)} \| T_{AA}(s) - s \|_1$

2. $\mathcal{L}_{cycle} = \sum_{s \in P(x)} \| T_{BA}(T_{AB}(s)) - s \|_1$
Phase II

1. \( \mathcal{L}_{REC_A} = \sum_{s \in P(x)} \| T_{AA}(s) - s \|_1 \)

2. \( \mathcal{L}_{cycle} = \sum_{s \in P(x)} \| T_{BA}(T_{AB}(s)) - s \|_1 \)

3. \( \mathcal{L}_{GAN_{AB}} = \sum_{s \in P(x)} -\ell(D_B(T_{AB}(s)), 0) \)

   \( \mathcal{L}_{D_{AB}} = \sum_{s \in P(x)} +\ell(D_B(T_{AB}(s)), 0) + \ell(D_B(s), 1) \)
Selective Backpropagation

When training our network with $x$ and its augmentations, backpropagation is applied selectively on the separate encoders and decoders only.

\[
\begin{align*}
T_{BB} &= G_B^U(G_S^S(E^S(E_B^U(x)))) \\
T_{BA} &= G_A^U(G_S^S(E^S(E_B^U(x)))) \\
T_{AA} &= G_A^U(G_S^S(E^S(E_A^U(x)))) \\
T_{AB} &= G_B^U(G_S^S(E^S(E_A^U(x))))
\end{align*}
\]
Selective Backpropagation

- Updating the shared encoder (Es) and decoder (Gs) with selective backpropagation turned off leads to **overfitting** on x, since for every shared representation, the unshared layers in domain A can still reconstruct this one sample.
Selective Backpropagation

- Updating the shared encoder (Es) and decoder (Gs) with selective backpropagation turned off leads to **overfitting** on x, since for every shared representation, the unshared layers in domain A can still reconstruct this one sample.
- However, as the shared encoder (Es) and decoder (Gs) can be trained with domain B samples **only**, translation from domain A to B is still possible.
Selective Backpropagation

• Updating the shared encoder (Es) and decoder (Gs) with selective backpropagation turned off leads to overfitting on x, since for every shared representation, the unshared layers in domain A can still reconstruct this one sample.

• However, as the shared encoder (Es) and decoder (Gs) can be trained with domain B samples only, translation from domain A to B is still possible.

• Use of a Patch GAN as well as convolutional layers induces further prior on the network that allows for successful translation given one input from domain A.
Qualitative Results
Qualitative Results
Figure 3: (a) Translating MNIST images to SVHN images. x-axis is the number of samples in $A$ (log-scale), y-axis is the accuracy of a pretrained classifier on the resulting translated images. The accuracy is averaged over 1000 independent runs for different samples. Blue: Our OST method. Yellow: UNIT [7]. Red: CycleGAN [2]. (b) The same graph in the reverse direction.
# Quantitative Results

Table 1: Ablation study for the MNIST to SVHN translation (and vice versa). We consider the contribution of various parts of our method on the accuracy. Translation is done for one sample.

<table>
<thead>
<tr>
<th>Augmentation</th>
<th>One-way cycle</th>
<th>Selective backprop</th>
<th>Accuracy (MNIST to SVHN)</th>
<th>Accuracy (SVHN to MNIST)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False</td>
<td>False</td>
<td>False</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td>True</td>
<td>False</td>
<td>False</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>False</td>
<td>True</td>
<td>False</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>True</td>
<td>True</td>
<td>False</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>False</td>
<td>True</td>
<td>True</td>
<td>0.19</td>
<td>0.20</td>
</tr>
<tr>
<td>True</td>
<td>False</td>
<td>True</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>False</td>
<td>True</td>
<td>True</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td>True</td>
<td>True</td>
<td>No Phase II update of $E^S$ and $G^S$</td>
<td>0.16</td>
<td>0.15</td>
</tr>
</tbody>
</table>

For two-way cycle:

| True | Two-way cycle | True | 0.20 | 0.13 |
| True | Two-way cycle | False | 0.11 | 0.12 |

| True | True | True | **0.23** | **0.23** |
Table 2: (i) Measuring the perceptual distance [29], between inputs and their corresponding output images of different style transfer tasks. Low perceptual loss indicates that much of the high-level content is preserved in the translation. (ii) Measuring the style difference between translated images and images from the target domain. We compute the average Gram matrix of translated images and images from the target domain and find the average distance between them, as described in [29].

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Content</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer2Winter</td>
<td>0.64</td>
<td>3.20</td>
<td>3.53</td>
<td>1.41</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>Winter2Summer</td>
<td>0.73</td>
<td>3.10</td>
<td>3.48</td>
<td>1.38</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Monet2Photo</td>
<td>3.75</td>
<td>6.82</td>
<td>5.80</td>
<td>1.46</td>
<td>1.41</td>
<td></td>
</tr>
<tr>
<td>Photo2Monet</td>
<td>1.47</td>
<td>2.92</td>
<td>2.98</td>
<td>2.01</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td>(ii) Style</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer2Winter</td>
<td>1.64</td>
<td>6.51</td>
<td>1.62</td>
<td>1.69</td>
<td>1.69</td>
<td></td>
</tr>
<tr>
<td>Winter2Summer</td>
<td>1.58</td>
<td>6.80</td>
<td>1.31</td>
<td>1.69</td>
<td>1.66</td>
<td></td>
</tr>
<tr>
<td>Monet2Photo</td>
<td>1.20</td>
<td>6.83</td>
<td>0.90</td>
<td>1.21</td>
<td>1.18</td>
<td></td>
</tr>
<tr>
<td>Photo2Monet</td>
<td>1.95</td>
<td>7.53</td>
<td>1.91</td>
<td>2.12</td>
<td>1.88</td>
<td></td>
</tr>
</tbody>
</table>
Quantitative Results

Table 3: (i) Perceptual distance [29] between the inputs and corresponding output images, for various drawing tasks. (ii) Style difference between translated images and images from the target domain. (iii) Correctness of translation as evaluated by a user study.

<table>
<thead>
<tr>
<th>Method</th>
<th>Images to Facades</th>
<th>Facades to Images</th>
<th>Images to Maps</th>
<th>Maps to Images</th>
<th>Labels to Cityscapes</th>
<th>Cityscapes to Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) OST 1</td>
<td>4.76</td>
<td>5.05</td>
<td>2.49</td>
<td>2.36</td>
<td>3.34</td>
<td>2.39</td>
</tr>
<tr>
<td>UNIT [7] All</td>
<td>3.85</td>
<td>4.80</td>
<td>2.42</td>
<td>2.30</td>
<td>2.61</td>
<td>2.18</td>
</tr>
<tr>
<td>CycleGAN [2] All</td>
<td>3.79</td>
<td>4.49</td>
<td>2.49</td>
<td>2.11</td>
<td>2.73</td>
<td>2.28</td>
</tr>
<tr>
<td>(ii) OST 1</td>
<td>3.57</td>
<td>7.88</td>
<td>2.24</td>
<td>1.50</td>
<td>0.67</td>
<td>1.13</td>
</tr>
<tr>
<td>UNIT [7] All</td>
<td>3.92</td>
<td>7.42</td>
<td>2.56</td>
<td>1.59</td>
<td>0.69</td>
<td>1.21</td>
</tr>
<tr>
<td>CycleGAN [2] All</td>
<td>3.81</td>
<td>7.03</td>
<td>2.33</td>
<td>1.30</td>
<td>0.77</td>
<td>1.22</td>
</tr>
<tr>
<td>(iii) OST 1</td>
<td>91%</td>
<td>90%</td>
<td>83%</td>
<td>67%</td>
<td>66%</td>
<td>56%</td>
</tr>
<tr>
<td>UNIT [7] ALL</td>
<td>86%</td>
<td>83%</td>
<td>81%</td>
<td>75%</td>
<td>63%</td>
<td>37%</td>
</tr>
<tr>
<td>CycleGAN [2] ALL</td>
<td>93%</td>
<td>84%</td>
<td>97%</td>
<td>81%</td>
<td>72%</td>
<td>45%</td>
</tr>
</tbody>
</table>
Future research

• One Shot Domain Adaptation
• One Shot Image to Image translation in the reverse direction
• Other Domains: Audio, Video?
• Online Setting
Thank You! Questions?
Minimality

• Potentially infinitely many solutions preserving distance correlations

Figure 1: An illustrative example where the two domains are line segments in $\mathbb{R}^2$. There are infinitely many mappings that preserve the uniform distribution on the two segments. However, only two stand out as “semantic”. These are exactly the two mappings that can be captured by a neural network with only two hidden neurons and Leaky ReLU activations, i.e., by a function $h(x) = \sigma_a(Wx + b)$, for a weight matrix $W$ and the bias vector $b$. 