CS 562
Presentation

Kung-Hsiang (Steeve) Huang
Papers to discuss

- Conditional Generative Adversarial Nets
- Video-to-Video Synthesis
Conditional Generative Adversarial Nets

Mehdi Mirza, Simon Osindero
Recap: Unconditional GAN

Unconditional GANs convert low-dimensional vectors into high-dimensional vectors. The objective function for such GANs is given by:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))].$$
Recap: Unconditional GAN

What if we want to control the attributes/properties of the characters generated?

\[
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))].
\]
Conditional GAN

$y$: Blonde hair

Low-dimensional vectors

$z$

High-dimensional vectors

$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}}[\log D(x|y)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z|y)))$
Training a conditional GAN

$y$: Blonde hair

$z$: Simple distribution

$G$:

$x = G(z|y)$

$D$:

Scalor: Is $x$ real? Does $x$, $y$ match?

$x$, Blonde hair, label: 0

$y$, Blonde hair, label: 1
Quantitative results

<table>
<thead>
<tr>
<th>Model</th>
<th>MNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBN [1]</td>
<td>138 ± 2</td>
</tr>
<tr>
<td>Stacked CAE [1]</td>
<td>121 ± 1.6</td>
</tr>
<tr>
<td>Deep GSN [2]</td>
<td>214 ± 1.1</td>
</tr>
<tr>
<td>Adversarial nets</td>
<td>225 ± 2</td>
</tr>
<tr>
<td>Conditional adversarial nets</td>
<td>132 ± 1.8</td>
</tr>
</tbody>
</table>

Table 1: Parzen window-based log-likelihood estimates for MNIST. We followed the same procedure as [8] for computing these values.
Qualitative results

Generated MNIST digits
**Limitation**

$y$: Blonde hair

$z$: Simple distribution

$y$: Blonde hair

$D$: Scaler: Is $x$ real? Does $x$, $y$ match?

$G$: $x = G(z|y)$

Fake image

(, Blonde hair, label: 0)

(, Blonde hair, label: 1)
Limitation

The generator and discriminator might not check if $x, y$ match.

$y$: Blonde hair

$z$: Simple distribution

$G$:

$x = G(z | y)$

Scaler: Is $x$ real? Does $x, y$ match?

$D$:

Fake image

(, Blonde hair, label: 1)

(, Blonde hair, label: 0)
Limitation

$y$: Blonde hair

$z$: Simple distribution

$G$: $x = G(z|y)$

Scaler: Is $x$ real? Does $x, y$ match?

($\cdot$, Blonde hair, label: 1)

($\cdot$, Blonde hair, label: 0)

Fake image
Applying (Conditional) GAN to NLP

● Issue: Textual outputs are discrete and non-differentiable.
● Solutions:
  ○ REINFORCE algorithm [1]
  ○ Gumbel-Softmax relaxation [2]

Takeaway

- The paper presents a simple yet effective extension towards controllability of GAN by feeding additional information to it.
- There is still a great room for improvement for synthesizing images using conditional GAN.
Video-to-Video Synthesis

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan Catanzaro
Video-to-video synthesis

Main challenge: Temporal coherence
Problem formulation

The model is tasked to map from a given a sequence of video frames

\[ s^T_1 \equiv \{s_1, s_2, ..., s_T \} \]

to a target sequence of video frames

\[ \tilde{x}^T_1 \equiv \{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_T \} \]

such that

\[ p(\tilde{x}^T_1 | s^T_1) = p(x^T_1 | s^T_1) \]
Problem formulation

The model is tasked to map from a given a sequence of video frames

$$s_1^T \equiv \{s_1, s_2, ..., s_T\}$$

Segmentation mask

to a target sequence of video frames

$$\tilde{x}_1^T \equiv \{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_T\}$$

such that

$$p(\tilde{x}_1^T | s_1^T) = p(x_1^T | s_1^T)$$

Real video clip
Optical flow

Approach

- Sequential generator: generate future frames
- Conditional image discriminator: ensure each frame is photorealistic
- Conditional video discriminator: ensure temporal consistency
Sequential generator

- Markov assumption

\[ p(\tilde{x}_1^T | s_1^T) = \prod_{t=1}^{T} p(\tilde{x}_t | \tilde{x}_{t-L}^{t-1}, s_{t-L}^t) \]

- Use a feed-forward network $F$ to approximate $p(\tilde{x}_t | \tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$

\[ F(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t) = (1 - \tilde{m}_t) \odot \tilde{w}_{t-1}(\tilde{x}_{t-1}) + \tilde{m}_t \odot \tilde{h}_t \]

- Estimated optical flow $\tilde{w}_{t-1} = W(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$
- Hallucinated image $\tilde{h}_t = H(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$
- Occlusion mask $\tilde{m}_t = M(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$
Conditional image discriminator

- Distinguish true pairs \((x_t, s_t)\) from fake \((\tilde{x}_t, s_t)\).
- Objective function:

\[
\mathcal{L}_I = E_{\phi_I(x^T_1, s^T_1)}[\log D_I(x_i, s_i)] + E_{\phi_I(\tilde{x}^T_1, s^T_1)}[\log(1 - D_I(\tilde{x}_i, s_i))].
\]
Conditional video discriminator

- Enhance temporal consistency by ensuring consecutive output frames resemble that of real frames **given the gold optical flow**.
- Distinguish true pairs \((x_{t-K}^{t-1}, w_{t-K}^{t-2})\) from fake \((\tilde{x}_{t-K}^{t-1}, w_{t-K}^{t-2})\).
- Objective function:

\[
\mathcal{L}_V = E_{\phi_V}(w_1^{T-1}, x_1^T, s_1^T) \left[ \log D_V(x_{i-K}^{i-1}, w_{i-K}^{i-2}) \right] \\
+ E_{\phi_V}(w_1^{T-1}, \tilde{x}_1^T, s_1^T) \left[ \log(1 - D_V(\tilde{x}_{i-K}^{i-1}, w_{i-K}^{i-2})) \right]
\]
Final objective function

\[
\min_F \left( \max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \mathcal{L}_V(F, D_V) \right) + \lambda_W \mathcal{L}_W(F)
\]

\[
\mathcal{L}_W = \frac{1}{T-1} \sum_{t=1}^{T-1} \left( \|\tilde{w}_t - w_t\|_1 + \|\tilde{w}_t(x_t) - x_{t+1}\|_1 \right)
\]
## Quantitative Results

Table 1: Comparison between competing video-to-video synthesis approaches on Cityscapes.

<table>
<thead>
<tr>
<th>Fréchet Inception Dist.</th>
<th>I3D</th>
<th>ResNeXt</th>
<th>Human Preference Score</th>
<th>short seq.</th>
<th>long seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>pix2pixHD</td>
<td>5.57</td>
<td>0.18</td>
<td>vid2vid (ours) / pix2pixHD</td>
<td><strong>0.87</strong> / 0.13</td>
<td><strong>0.83</strong> / 0.17</td>
</tr>
<tr>
<td>COVST</td>
<td>5.55</td>
<td>0.18</td>
<td>vid2vid (ours) / COVST</td>
<td><strong>0.84</strong> / 0.16</td>
<td><strong>0.80</strong> / 0.20</td>
</tr>
<tr>
<td>vid2vid (ours)</td>
<td><strong>4.66</strong></td>
<td><strong>0.15</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Quantitative Results

Table 2: Ablation study. We compare the proposed approach to its three variants.

<table>
<thead>
<tr>
<th>Human Preference Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>vid2vid (ours) / no background-foreground prior</td>
</tr>
<tr>
<td>vid2vid (ours) / no conditional video discriminator</td>
</tr>
<tr>
<td>vid2vid (ours) / no flow warping</td>
</tr>
</tbody>
</table>
## Quantitative Results

Table 3: Comparison between future video prediction methods on Cityscapes.

<table>
<thead>
<tr>
<th>Fréchet Inception Dist.</th>
<th>I3D</th>
<th>ResNeXt</th>
</tr>
</thead>
<tbody>
<tr>
<td>PredNet</td>
<td>11.18</td>
<td>0.59</td>
</tr>
<tr>
<td>MCNet</td>
<td>10.00</td>
<td>0.43</td>
</tr>
<tr>
<td>vid2vid (ours)</td>
<td><strong>3.44</strong></td>
<td><strong>0.18</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Human Preference Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>vid2vid (ours) / PredNet</td>
</tr>
<tr>
<td>vid2vid (ours) / MCNet</td>
</tr>
</tbody>
</table>
Qualitative Results

Figure 2: Apolloscape results. Left: pix2pixHD. Center: C0VST. Right: proposed. The input semantic segmentation mask video is shown in the left video. *Click the image to play the video clip in a browser.*
Why not regular seq2seq models?
A toy example

A toy example

Takeaway

- Enforcing temporal consistency is key to high quality video synthesis.
- By adding a video conditional discriminator, the work successfully shows improvement in temporal consistency of the videos generated.
- The proposed approach can be applied to various tasks, such as future video prediction.
References