Motivation

Goal:
- Model scene dynamics
- Make use of unlabeled video data

Why:
- Useful for video understanding and simulation

Approach:
- Two stream Generative Adversarial Network (GAN)
Generated Videos of Golf Course
Generated Videos of Beach
Generated Videos of Train Station
Generated Videos of (creepy) Baby
Future Generation
Background

- GANS
- Fractional Strided Convolution
- Spatio-temporal Convolution
- Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
  [Adford, Metz, Chintala 2016]
Background

- GANS
- Fractional Strided Convolution
- Spatio-temporal Convolution
- Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
  [Adford, Metz, Chintala 2016]
Generative Adversarial Networks
Generative Adversarial Networks
Generative Adversarial Networks

Two-player minimax game with value function

$$\min_{w_G} \max_{w_D} \mathbb{E}_{x \sim P(x)}[\log D(x; w_D)] + \mathbb{E}_{z \sim P(z)}[\log(1 - D(G(z; w_G); w_D))]$$
Minibatch SGD training of GANs

Alternating optimization

Update discriminator by gradient ascent

\[ \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(x^{(i)}) + \log(1 - D(G(z^{(i)})))] \]

Update generator by gradient descent

\[ \nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} [\log(1 - D(G(z^{(i)})))] \]
Background

- GANS
- **Fractional Strided Convolution**
- **Spatio-temporal Convolution**
- **Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks**
  [Adford, Metz, Chintala 2016]
Fractional Strided Convolution

- To stride by $1/n$ steps, pad $n-1$ zero blocks around each block and do unit striding
- Used to upsample
Background

- GANS
- Fractional Strided Convolution
- Spatio-temporal Convolution
- Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
  [Adford, Metz, Chintala 2016]
Spatio Temporal Convolution

- Analogous to normal 2D convolution over 2D or 3D space
- Using a 3D spatial convolution filter to stride over the 3D spatial as well as 4th temporal dimension
Background

- GANS
- Fractional Strided Convolution
- Spatio-temporal Convolution
- Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
  [Adford, Metz, Chintala 2016]
Combining GANs and CNNs

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
[Adford, Metz, Chintala 2016]
Combining GANs and CNNs

Lsun bedroom dataset

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
Generating Videos with Scene Dynamics
Generator-Discriminator Architecture

- One stream spatio-temporal generator
- Two stream foreground background generator
- Foreground-Background Mask
- Discriminator
Generator-Discriminator Architecture

- One stream spatio-temporal generator
- Two stream foreground background generator
- Foreground-Background Mask
- Discriminator
Generator-Discriminator Architecture: One Stream

[Diagram showing the architecture with 3D convolutions and a Tanh layer]

Noise (100 dim) → 3D convolutions → 64x64x2 (3) → Tanh

Generated Video
Space-Time Cuboid
Generator-Discriminator Architecture: One Stream

- Spatio-temporal provides invariance
- Fractionally stride up samples efficiently
Generator-Discriminator Architecture: One Stream

- Spatio-temporal provides invariance
- Fractionally stride up samples efficiently

Diagram:
- Noise (100 dim)
- Projected 4x4x2 Filter
- 3D convolutions
  - 64x64x32 (3)
- Tanh
- Generated Video
  - Space-Time Cuboid
Generator-Discriminator Architecture: One Stream

- Spatio-temporal provides invariance
- Fractionally stride up samples efficiently

Projected 4X4X2 Filter

Generated Video
Space-Time Cuboid

Fractional Striding 4X4X4 Filter
Generator-Discriminator Architecture

- One stream spatio-temporal generator
- Two stream foreground background generator
- Foreground-Background Mask
- Discriminator
Generator-Discriminator Architecture: Two Stream

Foreground Stream
3D convolutions

Background Stream
2D convolutions

Same structure as one stream for foreground

Noise
100 dim

Mask Sigmoid

Generated Video
Space-Time Cuboid

m ⊗ f + (1 − m) ⊗ b
Generator-Discriminator Architecture: Two Stream

- Foreground Stream: 3D convolutions
  - Same structure as one stream for foreground
  - Eliminates temporal dimension

- Background Stream: 2D convolutions

- Noise: 100 dim
Generator-Discriminator Architecture

- One stream spatio-temporal generator
- Two stream foreground background generator
- Foreground-Background Mask
- Discriminator
Generator-Discriminator Architecture: Mask

- **Foreground Stream**
  - 3D convolutions
  - Bounded in range [0, 1]
  - Gives weight to foreground vs background influence at each pixel

- **Background Stream**
  - 2D convolutions

- **Noise**
  - 100 dim

- **Mask**
  - Sigmoid
  - $m \odot f + (1 - m) \odot b$

- **Generated Video**
  - Space-Time Cuboid
  - Replicate over Time
Generator-Discriminator Architecture: Mask

- Mask matrix “cuts” out the foreground objects from the background, allowing for motion to occur for just the foreground.
- Sigmoid activation
Generator-Discriminator Architecture

- One stream spatio-temporal generator
- Two stream foreground background generator
- Foreground-Background Mask
- Discriminator
Generator-Discriminator Architecture: Discriminator

Recognizes realistic motion
Downsample instead of upsample

Video — Discriminator
3D convolutions

Real or Fake?
Generator-Discriminator Architecture: Discriminator

Discriminator
3D convolutions

Recognizes realistic motion
Downsample instead of upsample

Video ——

64x64x32 (3) —— 32x32x16 (64) —— 16x16x8 (128) —— 8x8x4 (256) —— 4x4x2 (512) —— 1x1x1 (2)

4X4X4 Filters
Normal Striding

Real or Fake?
Generator-Discriminator Architecture: Discriminator

- Recognizes realistic motion
- Downsamples instead of upsampling

3D convolutions

- Video
- 64x64x32 (3)
- 32x32x16 (64)
- 16x16x8 (128)
- 8x8x4 (256)
- 4x4x2 (512)
- 1x1x1 (2)

- 4x4x4 Filters
- Normal Striding
- 4x4x2 Filter
- Binary Classification

Real or Fake?
Training

- Intuition
- Dataset Acquisition
- Dataset Processing
Training

- Intuition
- Dataset Acquisition
- Dataset Processing
Training: Intuition

- Train both discriminator and generator in alternation.
- Discriminator wants to learn what makes a real image/video.
- Generator wants to fool discriminator by producing real looking image/video.

Generative adversarial networks (conceptual)
Training

- Intuition
- Dataset Acquisition
- Dataset Processing
Training: Dataset Acquisition

- 2M unlabelled videos from Flickr
- 5k hours of unfiltered video
- Rest filtered video with Places2
- 4 filtered categories of videos: golf course, babies, beaches, train station
- Normalizes to [-1, 1]
Training

- Intuition
- Dataset Acquisition
- **Dataset Processing**
Training: Dataset Processing

- Stabilizes data using SIFT and RANSAC to determine homography between frames
- Uses previous frame to fill in missing values, reject overly large changes
Experiments

- Action Classification
- Video Generation
- Future Generation
Experiments

- Action Classification
- Video Generation
- Future Generation
Action Classification

The Video GAN model learns a useful representation that can be transferred to Action Classification

- Train on 5k hours of unlabeled Flicker videos
- Fine-tune the discriminator on UCF101 action classification dataset
- Replace last layer of discriminator to be a K-way softmax classifier
Action Classification
The Video GAN model learns a useful representation that can be transferred to Action Classification

- Train on 5k hours of unlabeled Flicker videos
- Fine-tune the discriminator on UCF101 action classification dataset
- Replace last layer of discriminator to be a K-way softmax classifier

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>0.9%</td>
</tr>
<tr>
<td>STIP Features [36]</td>
<td>43.9%</td>
</tr>
<tr>
<td>Temporal Coherence [10]</td>
<td>45.4%</td>
</tr>
<tr>
<td>Shuffle and Learn [25]</td>
<td>50.2%</td>
</tr>
<tr>
<td>VGAN + Random Init</td>
<td>36.7%</td>
</tr>
<tr>
<td>VGAN + Logistic Reg</td>
<td>49.3%</td>
</tr>
<tr>
<td><strong>VGAN + Fine Tune</strong></td>
<td><strong>52.1%</strong></td>
</tr>
<tr>
<td>ImageNet Supervision [47]</td>
<td>91.4%</td>
</tr>
</tbody>
</table>
Action Classification

The Video GAN model learns a useful representation that can be transferred to Action Classification

- Train on 5k hours of unlabeled Flicker videos
- Fine-tune the discriminator on UCF101 action classification dataset
- Replace last layer of discriminator to be a K-way softmax classifier
Action Classification

The Video GAN model learns a useful representation that can be transferred to Action Classification

- Train on 5k hours of unlabeled Flicker videos
- Fine-tune the discriminator on UCF101 action classification dataset
- Replace last layer of discriminator to be a K-way softmax classifier
Experiments

- Action Classification
- Video Generation
- Future Generation
Generating Video

Video → Autoencoder
3D convolutions

Latent z code

64x64x32 (3) → 32x32x16 (64) → 16x16x8 (128) → 8x8x4 (256) → 4x4x2 (512) → 100D
Generating Video

- Fit a 256 component Gaussian Mixture Model over the autoencoder 100-dim outputs
- To generate video, sample from GMM a 100-dim latent code
- Latent code passed to generator to generate 32-frame video
Generating Video

100D vector sampled from GMM

Foreground Stream

3D convolutions

Background Stream

2D convolutions

Foreground

Tanh

Mask

Sigmoid

Generated Video

Space-Time Cuboid

Replicate over Time

$\mathbf{m} \odot \mathbf{f} + (1 - \mathbf{m}) \odot \mathbf{b}$
Generated Videos of Beach
Generated Videos of (creepy) Baby
Mechanical Turk Evaluation

- 13,000 opinions over 4 categories by 150 workers
- Mostly prefers VGAN over autoencoder
- Unexpected number prefers VGAN over real

<table>
<thead>
<tr>
<th>“Which video is more realistic?”</th>
<th>Percentage of Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Golf</td>
</tr>
<tr>
<td>Random Preference</td>
<td>50</td>
</tr>
<tr>
<td>Prefer VGAN Two Stream over Autoencoder</td>
<td>88</td>
</tr>
<tr>
<td>Prefer VGAN One Stream over Autoencoder</td>
<td>85</td>
</tr>
<tr>
<td>Prefer VGAN Two Stream over VGAN One Stream</td>
<td>55</td>
</tr>
<tr>
<td>Prefer VGAN Two Stream over Real</td>
<td>21</td>
</tr>
<tr>
<td>Prefer VGAN One Stream over Real</td>
<td>17</td>
</tr>
<tr>
<td>Prefer Autoencoder over Real</td>
<td>4</td>
</tr>
</tbody>
</table>
Experiments

- Action Classification
- Video Generation
- Future Generation
Future Generation
Future Generation
Summary: Motivation

- Model scene dynamics
- Make use of unlabeled video data
Summary: Architecture

- One stream spatial temporal generator
- Two stream foreground background generator
- Discriminator recognizes real and generated videos
Summary: Experimentation and Results

- VGAN successful classification of action
- Randomly generating video with z sample from GMM
- Generating future video with single frame
Questions?
Questions?
Questions?
Questions?
Questions?