Hierarchical Long-term Video Prediction without Supervision ICML 2018
Nevan Wichers, Ruben Villegas, Dumitru Erhan, Honglak Lee

Seminar: Computer Vision SS19
Stefan Möhrle
Introduction

- Given: $C$ context frames
- Goal: predict future frames

*Wichers et al. 2018*
Motivation

• Self-supervision
  ◦ Autonomously collect virtually unlimited experience
  ◦ Make huge amounts of data usable

• Video prediction
  ◦ Learn representations of motion/actions
  ◦ Future prediction for interacting agents

⇒ Unsupervised feature learning

⇒ Accurate model of the world for reinforcement learning
Related Work

Frame-level

- Predicts pixel movement
- Observes own predictions
Finn et al (2016)

- Predict the motions of different objects / image segments
- Merge predictions via masking
- Convolutional dynamic neural advection (CDNA)
  - 10 normalized transformation kernels
  - 11-channel composing mask

Finn et al. 2016
Denton and Fergus (2018)

- Deterministic frame predictor
- Time-dependent stochastic latent variables $z_t$
  - Modeled by a separate convolutional LSTM
  - High variance in periods of uncertainty

*Denton & Fergus 2018*
Related Work

Frame-level

- Predicts pixel movement
- Observes own predictions
Related Work

Frame-level

- Predicts pixel movement
- Observes own predictions

Hierarchical

- Predicts high-level features
- Separation between prediction and image generation
Villegas et al (2017)

- $p_t$: human pose landmarks, $C$: #Contextframes
Villegas et al (2017)

- Separate training of LSTM and decoder
Approach

- Based on network architecture of Villegas et al.
- No ground truth annotations required
- Joint training
Approach

- Use $I_1$ instead of $I_C$ for the decoder
Approach

- Use general feature vector with auto-regressive connection
Decoder

- Deep version visual analogy network (VAN, Reed et al., 2015)
  
  \[ \tilde{I}_t, M_t = \text{VAN}(e_1, \hat{e}_t, I_1) \]
  
  \[ \hat{I}_t = \tilde{I}_t \odot M_t + I_1 \odot (1 - M_t) \]
Decoder

- \( \text{VAN}(e_1, \hat{e}_t, I_1) = \)

\[
g_{\text{dec}}(f_{\text{enc}}(\hat{e}_t) + T(f_{\text{img}}(I_1), f_{\text{enc}}(e_1), f_{\text{enc}}(\hat{e}_t)))
\]

Reed et al. 2015
End to End

- Train end-to-end with L2 loss
- Produces blurry predictions
Encoder Predictor with Visual Analogy (EPVA)

- Train decoder without uncertainty of predictor
Encoder Predictor with Visual Analogy (EPVA)

- Train predictor without pixel-level noise
EPVA Adversarial Loss

- Reduce blurry predictions with discriminator network
Evaluation: Toy dataset

- Bouncing shape with changing size
- Context: 3 frames
- Train: 16 frames
- Evaluate: ~1000 frames
- Test: shape color and presents (frame 1012 to 1022)
Evaluation: Toy dataset

- Bouncing shape with changing size
- Context: 3 frames
- Train: 16 frames
- Evaluate: ~1000 frames
- Test: shape color and presents (frame 1012 to 1022)
Evaluation: Toy dataset

Wichers et al. 2018
Evaluation: Toy dataset

- Bouncing shape with changing size
- Context: 3 frames
- Train: 16 frames
- Evaluate: ~1000 frames
- Test: shape color and presents (frame 1012 to 1022)

<table>
<thead>
<tr>
<th>Method</th>
<th>Shape has correct color</th>
<th>Shape has wrong color</th>
<th>Shape disappeared</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPVA</td>
<td>96.9%</td>
<td>3.1%</td>
<td>0%</td>
</tr>
<tr>
<td>Finn et al. (2016)</td>
<td>24.6%</td>
<td>5.7%</td>
<td>69.7%</td>
</tr>
</tbody>
</table>
Evaluation: Human3.6M

- Subsampled to 6.25 frames per second
- Context: 5 frames (0.8s)
- Train: 32 frames
- Evaluate: 126 frames (~20s)
- Test: side by side comparison
Evaluation: Human3.6M

Wichers et al. 2018
Evaluation: Human3.6M
Evaluation: Human3.6M

- Subsampled to 6.25 frames per second
- Context: 5 frames
- Train: 32 frames
- Evaluate: 126 frames
- Test: side by side comparison

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Frames</th>
<th>Ours is better</th>
<th>Same</th>
<th>Baseline is better</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPVA vs Finn et al. (2016)</td>
<td>1-127</td>
<td>46.4%</td>
<td>40.7%</td>
<td>12.9%</td>
</tr>
<tr>
<td>EPVA Adv. vs Finn et al. (2016)</td>
<td>1-127</td>
<td>73.9%</td>
<td>13.2%</td>
<td>12.9%</td>
</tr>
<tr>
<td>EPVA Adv. vs Denton &amp; Fergus (2018)</td>
<td>5-127</td>
<td>58.2%</td>
<td>24.0%</td>
<td>17.8%</td>
</tr>
</tbody>
</table>
**Evaluation: Human3.6M**

- Subsampled to 6.25 frames per second
- Context: 5 frames
- Train: 32 frames
- Evaluate: 126 frames
- Test: side by side comparison

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Frames</th>
<th>Ours is better</th>
<th>Same</th>
<th>Baseline is better</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPVA vs Finn et al. (2016)</td>
<td>1-127</td>
<td>46.4%</td>
<td>40.7%</td>
<td>12.9%</td>
</tr>
<tr>
<td>EPVA Adv. vs Finn et al. (2016)</td>
<td>1-127</td>
<td>73.9%</td>
<td>13.2%</td>
<td>12.9%</td>
</tr>
<tr>
<td>EPVA Adv. vs Denton &amp; Fergus (2018)</td>
<td>5-127</td>
<td>58.2%</td>
<td>24.0%</td>
<td>17.8%</td>
</tr>
<tr>
<td>EPVA Adv. vs Finn et al. (2016)</td>
<td>63-127/1-63</td>
<td>67.2%</td>
<td>17.5%</td>
<td>15.3%</td>
</tr>
</tbody>
</table>
Evaluation: Human3.6M

Person detector

Person Detector Confidence for different methods

Frame number

Wichers et al. 2018
Evaluation: Feature transfer

- Learned features + 2-layer MLP
- Baseline: features by VVG-net trained for recognition
- Test: Output human pose landmarks
- 9% relative improvement

<table>
<thead>
<tr>
<th>Used features</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-net</td>
<td>0.0758</td>
</tr>
<tr>
<td>Ours</td>
<td>0.0687</td>
</tr>
</tbody>
</table>
Ablation Studies

Wichers et al. 2018
Conclusion

We saw:

- an unsupervised approach for discovering high-level features
- a joint training strategy
- long-term pixel-level video prediction for about 20 seconds
Hierarchical Long-term Video Prediction without Supervision ICML 2018
Nevan Wichers, Ruben Villegas, Dumitru Erhan, Honglak Lee

Seminar: Computer Vision SS19
Stefan Möhrle
References


