Unsupervised Learning of Video Representations using LSTMs

Nitish Srivastava
Elman Mansimov
Ruslan Salakhutdinov

ICML 2015
How to represent a video

- Model needs to capture both spatial and temporal information to learn good representations

- Supervised Learning: Action Recognition, Detection, etc.

- Does credit assignment (backprop) in a very high dimensional space. Needs lots of labelled data

- Unsupervised Learning: Autoencoding, Future Prediction, etc.
Sequence-to-sequence learning


\[
P(\text{target seq} \mid \text{input seq})
\]
Autoencoder

Encoder

Decoder

\[ v_1 \xrightarrow{W_1} v_2 \xrightarrow{W_1} \text{copy} \xrightarrow{W_2} \hat{v}_1 \xrightarrow{W_2} \hat{v}_2 \xrightarrow{W_2} \hat{v}_3 \]
## Comparison

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future Predictor</td>
<td>Short term information is more important to predict next frame.</td>
</tr>
<tr>
<td>Cannot cheat by copying</td>
<td></td>
</tr>
<tr>
<td>Autoencoder</td>
<td>High capacity auto encoder can cheat by memorizing inputs</td>
</tr>
<tr>
<td>Needs to remember long-term information</td>
<td></td>
</tr>
</tbody>
</table>

**Diagram:**

- **Future Predictor**:
  - $\hat{v}_4$
  - $\hat{v}_5$
  - $\hat{v}_6$

- **Autoencoder**:
  - $v_1$
  - $v_2$
  - $v_3$
  - $\hat{v}_3$
  - $\hat{v}_2$
  - $\hat{v}_1$
Composite Model

Encoder

\( W_1 \)

\( \hat{v}_3 \)

\( \hat{v}_2 \)

\( \hat{v}_1 \)

\( W_2 \)

\( \hat{v}_4 \)

\( \hat{v}_5 \)

\( \hat{v}_6 \)

\( W_3 \)

\( \hat{v}_3 \)

\( \hat{v}_2 \)

\( \hat{v}_1 \)

copy
To condition or not to condition?

- If the target distribution has multiple modes, a conditional decoder will help capture them.

- Future predictor - multimodal

- During training - condition on the true frame.

- At test time - use generated frame from previous step without noise

- Recent work by S. Bengio et al 2015 explores this problem more in detail
Datasets

Moving MNIST

32x32 Patches

Percepts (from a Conv Net)

features
Results

Input sequence → Ground truth future →

Input reconstruction ← Future prediction ←

One layer Composite Model
Two layer Composite Model
Two layer Conditional Composite Model
Future prediction over longer time scales

Model trained to predict 10 frames into the future
Running for 100 frames into the future
Future prediction over longer time scales

Model trained to predict 10 frames into the future
Running for 100 frames into the future
Out of domain data

Model trained on two moving digits

Testing on one moving digit

Testing on three moving digits
Results on Patches

Input Reconstruction  Future Prediction  Input Reconstruction  Future Prediction
Action Recognition

• Supervised:
  
  • UCF-101 (Soomro et al. 2012) 13,320 videos (9.5K training), avg length 6.2s.
  
  • HMDB-51 (Kuehne et al. 2011) 5100 videos (3570 training), avg length 3.2s

• Unsupervised: Subset of Sports-1M (Karpathy et al. 2014), 10s clips randomly chosen, total 300 hours.

• Extract percepts by pushing video frames through a good convolutional net trained on ImageNet (Simonyan et al. 2014).
Action Recognition

- LSTM Classifier

- Try to predict the action at each step.

- Applied on sequences of 16 frames. Average predictions made at each step.

- Average 16 frame chunks with a stride of 8 over the video.

- Similar experiments on 4 optical flow frames with stride of 1 over the video
## Action Recognition Results

<table>
<thead>
<tr>
<th>Model</th>
<th>UCF-101 RGB</th>
<th>UCF-101 Optical Flow</th>
<th>HMDB-51 RGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Frame</td>
<td>72.2</td>
<td>72.2</td>
<td>40.1</td>
</tr>
<tr>
<td>LSTM Classifier (Random initialization)</td>
<td>74.5</td>
<td>74.3</td>
<td>42.8</td>
</tr>
<tr>
<td>LSTM Classifier (Pre-trained Composite model Encoder)</td>
<td>75.8</td>
<td>74.9</td>
<td>44.1</td>
</tr>
</tbody>
</table>
Action Recognition Results

More improvement for small labelled datasets

UCF-101

HMDB-51
# Action Recognition Results

<table>
<thead>
<tr>
<th>Model</th>
<th>UCF-101 small</th>
<th>UCF-101</th>
<th>HMDB-51 small</th>
<th>HMDB-51</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline LSTM</td>
<td>63.7</td>
<td>74.5</td>
<td>25.3</td>
<td>42.8</td>
</tr>
<tr>
<td>Autoencoder</td>
<td>66.2</td>
<td>75.1</td>
<td>28.6</td>
<td>44.0</td>
</tr>
<tr>
<td>Future Predictor</td>
<td>64.9</td>
<td>74.9</td>
<td>27.3</td>
<td>43.1</td>
</tr>
<tr>
<td>Conditional Autoencoder</td>
<td>65.8</td>
<td>74.8</td>
<td>27.9</td>
<td>43.1</td>
</tr>
<tr>
<td>Conditional Future Predictor</td>
<td>65.1</td>
<td>74.9</td>
<td>27.4</td>
<td>43.4</td>
</tr>
<tr>
<td>Composite Model</td>
<td>67.0</td>
<td>75.8</td>
<td>29.1</td>
<td>44.1</td>
</tr>
<tr>
<td>Composite Model with a Conditional Future Predictor</td>
<td>67.1</td>
<td>75.8</td>
<td>29.2</td>
<td>44.0</td>
</tr>
</tbody>
</table>

UCF-101 small: 10 videos per class  
HMDB-51 small: 4 videos per class
Action Recognition Results

• The best action recognition results are still quite far away.

• IDT + Multi-skip feature stacking (Lan et al 2014) : 89.1%

• Two stream convolutional nets (Simonyan et al 2014) : 88.0%

• C3D (3 nets) (Tran et al 2014) + iDT (Wang et al 2013): 90.4%

• Our best result (averaging LSTM models on flow and RGB) : 84.3%
Conclusions

• Explored different **encoder-decoder** models that learn video representations in **unsupervised** manner.

• Best Performance: **Composite Model**.

• On Moving MNIST, the model was able to **persistently generate motion** well beyond the time scales it was trained for, but lost precise digit identity.

• **Main Flow**: Lack of Stochasticity in the Decoder.

• **In future**, we hope that unsupervised learning will give larger improvement on supervised tasks.
Thanks :)