Generative and Predictive Models of Videos for Understanding the World

Oleh Rybkin

(some slides taken from Drew Jaegle, Karl Pertsch)
Can predictive objectives be useful for semantic understanding?

- Objects?
- Events?
- Affordances?

Credits: (left) Francis Vachon, found in a Chelsea Finn presentation, (right) Sergey Levine
Learning what you can do before doing anything

Oleh Rybkin*, Karl Pertsch*, Konstantinos G. Derpanis, Kostas Daniilidis, Andrew Jaegle

ICLR 2019
Understanding actions

Credit: found via Nando De Freitas tweets
Understanding actions

$z_1 = \text{RIGHT}$

$z_2 = \text{UP}$

$z_3 = g(z_1, z_2) = \text{RIGHT} + \text{UP}$

Learning what you can do before doing anything. ICLR 2019.
Variational Video Prediction

\[ \mathcal{L}_2 \]

\( \mathcal{N}(0, I) \) - \( z_t \)

KL

Prior - Posterior

\( LSTM \)

\( \hat{x}_t \)

\( CNN_e \)

\( x_{t-1} \) - \( x_t \)

Sampling

Loss

Denton & Fergus, 2018; Lee et al., 2018. Chung et al., 2015

Learning what you can do before doing anything. ICLR 2019.
Variational Video Prediction

Learning what you can do before doing anything. ICLR 2019.
Variational Video Prediction with Information Bottleneck

The (beta-)VAE objective for stochastic video prediction is:

$$\sum_t \left[ \mathbb{E}_{q(z_t|x_{t-1:t})} \log p(x_t|Z_t, x_{t-1}) - \beta \text{KL}[q(Z_t|x_{t-1:t}), p(Z)] \right]$$

Which is equivalent to the VIB lower bound of the following:

$$\max I((z_t, x_{t-1}); x_t) \text{ s.t. } I(z_t; x_{t-1:t}) \leq I_c.$$
Enforcing structure with composability

\[ z_1 = \text{RIGHT} \]

\[ z_2 = \text{UP} \]

\[ z_3 = g(z_1, z_2) = \text{RIGHT} + \text{UP} \]
CLASP: Enforcing structure with composability

$\mathcal{L}_2$

$z_2 \rightarrow z_3 \rightarrow z_4$

$\nu_1 \rightarrow \mathcal{N}(0, I)$

$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4$

$\text{CNN}_{\text{c}} \rightarrow \text{MLP}_{\text{inf}} \rightarrow \text{MLP}_{\text{comp}} \rightarrow \text{LSTM} \rightarrow \text{CNN}_{\text{c}}$
Reacher environment
Understanding actions

a) CLASP (Ours)

b) Denton & Fergus (2018)

Learning what you can do before doing anything. ICLR 2019.
Applications of CLASP

Passive learning

Input Videos (no actions)

Video Predictions

Action-Free Video Prediction Training

Active learning

Input Videos (with actions $u_t$)

Few-Shot Latent-Action Bijection Learning

Input Image & Action Sequence

Action-Conditioned Video Prediction

Start & Target Image

Planning in representation space

Learning what you can do before doing anything. ICLR 2019.
Action-conditioned prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>Reacher [deg]</th>
<th>BAIR [px]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>26.6 ± 21.5</td>
<td>-</td>
</tr>
<tr>
<td>Baseline</td>
<td>22.6 ± 17.7</td>
<td>3.6 ± 4.0</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>2.9 ± 2.1</strong></td>
<td><strong>3.0 ± 2.1</strong></td>
</tr>
<tr>
<td>Supervised</td>
<td>2.6 ± 1.8</td>
<td>2.0 ± 1.3</td>
</tr>
</tbody>
</table>

Learning what you can do before doing anything. ICLR 2019.
Applications of CLASP

Passive learning

Active learning

Input Videos (no actions)

Video Predictions

Action-Free Video Prediction Training

Input Images & Action Sequence

Action-Conditioned Video Prediction

Start & Target Image

Planning in representation space

Few-Shot Latent-Action Bijection Learning

Ebert et Finn et al., 2018

Learning what you can do before doing anything. ICLR 2019.
Planning in learned latent space

Learning what you can do before doing anything. ICLR 2019.
Varying visual characteristics

Learning what you can do before doing anything. ICLR 2019.
1. The *inductive biases* of minimality and composability provide sufficient constraints for learning action representations just from visual observations.

2. The learned representation is *disentangled* from the static scene content and visual characteristics of the environment.

3. The representation to be used for *planning* and *action-conditioned prediction* while requiring orders of magnitude less action-labeled videos.
KeyIn: Discovering Subgoal Structure with Keyframe-based Video Prediction

Karl Pertsch*, Oleh Rybkin*, Jingyun Yang, Konstantinos G. Derpanis, Joseph Lim, Kostas Daniilidis, Andrew Jaegle
Dynamics in complex scenes are stochastic. But not uniformly so!
How can we exploit this structure to improve long-term reasoning?
**Keyframes**: capture interesting structure in time, but also allow reconstruction of the full dynamics.
Keyframing

- Discover keyframes
- Improve long-term reasoning

1. Draw the start and end points of all motions: define the stochastic long-term sequence dynamics (*lead animator*).
2. Interpolate between the start and end points: make the local, deterministic dynamics explicit (*inbetweener*).
Keyln - keyframe prediction

LSTM_{key}

I_0 \rightarrow I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow I_4 \rightarrow I_5 \rightarrow I_6

Time

Frames

z_1 \rightarrow z_2
KeyIn - keyframe-based prediction

LSTM_{key}

Keyframes

LSTM_{inter}

Frames

Time

$K^0$ $K^1$ $K^2$

$I_0$ $I_1$ $I_2$ $I_3$ $I_4$ $I_5$ $I_6$

$z_1$ $z_2$
Keyln - predicting interframe offsets

LSTM_{key}

Keyframes

K^0

δ^1

K^1

δ^2

K^2

LSTM_{inter}

Frames

I_0

I_1

I_2

I_3

I_4

I_5

I_6

Time
KeyIn - Continuous relaxation

LSTM
c_{key}

Keyframes

Offset

Loss

Targets

\[ \delta^1 \]
Keyln - Full loss

\[ \mathcal{L}_{key} = \left( \sum_{t} c^t \beta_{ki} \right) \| \hat{K}^t - \tilde{K}^t \|^2 \]

- Soft Keyframe targets
- Soft embedding targets
- Prior divergence
- Interpolation targets
KeyIn - full method

- LSTM_key
- Keyframes
  - $K^0$
  - $K^1$
  - $K^2$
  - $\delta^1$
  - $\delta^2$
- LSTM_inter
- Frames
  - $I_0$
  - $I_1$
  - $I_2$
  - $I_3$
  - $I_4$
  - $I_5$
  - $I_6$
- Time
Structured Brownian motion data
Enforcing descriptive Keyframes
Generative model of trajectories via keyframes

Ground Truth

Predicted

Legend

Keyframes

Input

Predicted
Pushing data
Planning

Demonstration Data

Train on Video Prediction

Keyframe Model

Plan subgoals

Initial frame

Subgoal 1

Subgoal n

Target frame

Reach Subgoals with Control

a₁ ... aₖ

a₁ ... aₖ

Ebert et Finn et al., 2018
Algorithm 1 Planning in the subgoal space.

Input: Keyframe model $\text{KEYIN}(., .)$, cost function $c$
Input: Start and target images $I_0$ and $I_{\text{target}}$

Sample $L$ sequences of latent variables:
$$z^{0:M} \sim \mathcal{N}(\mu_n, \sigma_n)$$

Produce subgoal plans: $\hat{K}^{0:M} = \text{KEYIN}(I_0, z^{0:M})$

Compute cost between produced and true target:
$$c(\hat{K}^M)$$

Choose $L'$ best plans,

end for

Return: Best subgoal plan $K^{0:M}$
Planning on the pushing task

<table>
<thead>
<tr>
<th>Method</th>
<th>Final position error</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>1.32 ± 0.06</td>
<td>-</td>
</tr>
<tr>
<td>Random</td>
<td>1.32 ± 0.07</td>
<td>-</td>
</tr>
<tr>
<td>No Subgoals</td>
<td>0.90 ± 0.14</td>
<td>15.0%</td>
</tr>
<tr>
<td>TAP</td>
<td>0.80 ± 0.16</td>
<td>23.3%</td>
</tr>
<tr>
<td>Jumpy</td>
<td>0.62 ± 0.33</td>
<td>58.8%</td>
</tr>
<tr>
<td>KeyIn (Ours)</td>
<td>0.50 ± 0.26</td>
<td>64.2%</td>
</tr>
</tbody>
</table>
KeyIn: Discovering Subgoal Structure with Keyframe-based Video Prediction

- The model learns to predict videos by first predicting a set of descriptive keyframes.
- A differentiable loss allows to train the model to select the most descriptive keyframes.
- The keyframes the model discovers are useful as subgoals for a planning task.