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

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

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:

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

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

\[ z_1 = \text{RIGHT} \]

\[ z_2 = \text{UP} \]

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

Learning what you can do before doing anything. ICLR 2019.
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

Ground Truth:

CLASP (ours):

Denton & Fergus:

<table>
<thead>
<tr>
<th>Method</th>
<th>Reacher Error [deg]</th>
<th>BAIR Error [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>2.9 ± 2.1</td>
<td>3.0 ± 2.1</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

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

Action-Conditioned Video Prediction

Planning in representation space

Input Image & Action Sequence

Start & Target Image

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

In submission to ICML
Keyframes in natural sequences

- 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*).
KeyIn - keyframe prediction

LSTM_{\text{key}}

Frames

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

Time

z_1

z_2
KeyIn - keyframe-based prediction

LSTM\textsubscript{key}

Keyframes

LSTM\textsubscript{inter}

Frames

Time
KeyIn - predicting interframe offsets

LSTM\text{\tiny key} \quad LSTM\text{\tiny inter}

Keyframes

Frames

Time

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

K^0 \quad K^1 \quad K^2

\delta^1 \quad \delta^2

z_1 \quad z_2
KeyIn - Continuous relaxation

LSTM_{key} → Keyframes → Offset → Loss → Targets

\( K_1 \) → \( \hat{K}_1 \) → \( K_2 \)

\( I_0 \) → \( I_1 \) → \( I_2 \) → \( I_3 \) → \( I_4 \) → \( I_5 \) → \( I_6 \) → \( \delta_1 \)
KeyIn - Full loss

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

LSTM\textsubscript{key}

Keyframes

K\textsuperscript{0}

K\textsuperscript{1}

K\textsuperscript{2}

LSTM\textsubscript{inter}

Frames

I\textsubscript{0}

I\textsubscript{1}

I\textsubscript{2}

I\textsubscript{3}

I\textsubscript{4}

I\textsubscript{5}

I\textsubscript{6}

Time

\delta\textsuperscript{1}

\delta\textsuperscript{2}
Structured Brownian motion data
Enforcing descriptive Keyframes

Jumpy (Baseline)

Ground Truth

Predicted Keyframes

Predicted Image Sequence
Generative model of trajectories via keyframes

Ground Truth | Predicted

Legend:
- Input
- Keyframes
- t
- Predicted
Pushing data

Input
Keyframes
Frames

Input
Keyframes
Frames
Planning

Demonstration Data

Train on Video Prediction

Keyframe Model

Plan subgoals

Initial frame

Subgoal 1

Subgoal n

Target frame

Reach Subgoals with Control

$a_1$ ...

$a_k$

$a_1$ ...

$a_k$

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:
$$z^{0:M} \sim \mathcal{N}(\mu, \sigma)$$

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

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

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><strong>KEYIN (Ours)</strong></td>
<td><strong>0.50 ± 0.26</strong></td>
<td><strong>64.2%</strong></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.