Unsupervised Representations towards Counterfactual Predictions

Animesh Garg
Compositional Representations

Vacuuming  Sweeping/Mopping  Cooking  Laundry
Compositional Representations

Diversity:
New Scenes, Tools,…

Complexity:
Long-term Settings
Compositional Representations

Unstructured/Unknown New Environment

Dartmouth AI Meeting

1st Industrial robot

1956 61 1968

2013 2018 2020
Compositional Representations

Supervision → Task Imitation → Generalization

Input

Task Performance in Data Scarce Set-up

New Task Variations in Novel Environments
Compositional Representations

- Learning Disentanglement
- Learning Keypoints
- Learning Causal Graphs
Compositional Representations

Learning Disentanglement

Learning Keypoints

Learning Causal Graphs
Generative Models: Disentanglement

Objectives of Disentanglement
- Compositional Representations
- Controllable Sample Generation

dSprites 3DShapes

Existing datasets in unsupervised disentanglement learning
Disentanglement: Challenges

✗ High-Resolution Output

✗ Non-identifiability in Unsupervised setting

✗ Metrics focus on learning disentangled representations

Locatello et al. ICML 2019.
Disentanglement: Challenges

✗ High-Resolution Output
StyleGAN based backbone (~1%)

New high-resolution synthetic datasets: Falcor3D and Isaac3D

✗ Non-identifiability in Unsupervised setting

Limited Supervision (~1%)

✗ Metrics focus on learning disentangled representations

New Metric to Trade-off between controllability and disentanglement
Disentanglement: StyleGAN

- Used a style-based generator to replace traditional generator
- Success at generating high-resolution realistic images
The semi-supervised loss is given by

$$\mathcal{L}^{(G)} = \mathcal{L}_{GAN} + \gamma_G \mathcal{L}_{unsup} + \alpha \mathcal{L}_{sr}$$

$$\mathcal{L}^{(D,E)} = - \mathcal{L}_{GAN} + \gamma_E \mathcal{L}_{unsup} + \beta \mathcal{L}_{sup} + \alpha \mathcal{L}_{sr}$$

with

$$\mathcal{L}_{unsup} = \sum_{c \sim \mathcal{C}, z \sim p_z} \| E(G(c, z)) - c \|_2$$  \hspace{1cm} \text{unsupervised InfoGAN loss term}$$

$$\mathcal{L}_{sup} = \sum_{(x,c) \sim \mathcal{J}} \| E(x) - c \|_2$$  \hspace{1cm} \text{supervised label reconstruction term}$$

$$\mathcal{L}_{sr} = \sum_{(x,c) \sim \mathcal{M}} \| E(x) - c \|_2$$  \hspace{1cm} \text{smoothness regularization term}$$

**Disentanglement in StyleGAN**

**Mapping Network** in the generator conditions on the factor code and the encoder predicts its value
Disentanglement: Semi-StyleGAN

The semi-supervised loss is given by

$$\mathcal{L}^{(G)} = \mathcal{L}_{GAN} + \gamma_G \mathcal{L}_{unsup} + \alpha \mathcal{L}_{sr}$$

$$\mathcal{L}^{(D,E)} = - \mathcal{L}_{GAN} + \gamma_E \mathcal{L}_{unsup} + \beta \mathcal{L}_{sup} + \alpha \mathcal{L}_{sr}$$

with

$$\mathcal{L}_{unsup} = \sum_{c \sim \mathcal{C}, z \sim p_z} \| E(G(c, z)) - c \|_2$$  → unsupervised InfoGAN loss term

$$\mathcal{L}_{sup} = \sum_{(x,c) \sim \mathcal{J}} \| E(x) - c \|_2$$  → supervised label reconstruction term

$$\mathcal{L}_{sr} = \sum_{(x,c) \sim \mathcal{M}} \| E(x) - c \|_2$$  → smoothness regularization term

Disentanglement in StyleGAN Mapping Network in the generator conditions on the factor code and the encoder predicts its value.

Labeled Data $|\mathcal{J}| \ll |\mathcal{X}|$ Unpaired Data $\mathcal{M}$: Artificially Augmented Data

Li et al. ICML 2020.
With very limited supervision, Semi-StyleGAN can achieve good disentanglement on real data.
Semi-StyleGAN: Isaac3D (512x512)

Each factor in the interpolated images changes smoothly without affecting other factors.
Semi-StyleGAN: Falcor3D (512x512)

Each factor in the interpolated images changes smoothly without affecting other factors.
Semi-StyleGAN: Role of Limited Supervision

Only using 0.25∼2.5% of labeled data at par with supervised disentanglement

Semi-StyleGAN with the default setting  \( \gamma_G = \beta = \gamma, \gamma_E = 0, \alpha = 1 \)
Semi-StyleGAN: Fine-Grained Tuning

New architecture with same loss model for semantic fine-grained image editing
We randomly choose some deep learning researchers as test images.
Semi-StyleGAN: CelebA (256x256)

We shift the robot position to the right side, and attach it with an unseen object in test images.
Compositional Representations

- Learning Disentanglement
- Learning Keypoints
- Learning Causal Graphs
Representations for multi-step reasoning in Robotics under physical and semantic constraints
Model-based learning

choose action sequence

\[ a_t, \ldots, a_{t+H} \]

dynamics

\[ s' \sim f(\cdot | s, a) \]

[Deisenroth et al, RSS’07], [Guo et al, NeurIPS’14], [Watter et al, NeurIPS’15], [Finn et al, ICRA’17], ......
Model-based learning

[Deisenroth et al. RSS’07]  [Agrawal et al. ICRA’16]

[Ebert et al. CoRL’17]  [Janer et al. ICRA’19]
CAVIN: Hierarchical planning in learned latent spaces

Leverage **Hierarchical Abstraction** in Action Space
Without **Hierarchical Supervision**

CAVIN Planner

**effect code c**

**motion code z**
CAVIN: Hierarchical planning in learned latent spaces

- CAVIN Planner
- effect code c
- motion code z
- subgoals

Diagram showing a robotic arm and a grid with objects, illustrating the planning process.
CAVIN: Hierarchical planning in learned latent spaces

CAVIN Planner

effect code $c$

motion code $z$

actions
CAVIN: Hierarchical planning in learned latent spaces

choose $c \sim p(c)$

meta-dynamics
CAVIN: Hierarchical planning in learned latent spaces

Choose $c \sim p(c)$

$S_0$
Hierarchical planning in learned latent spaces

choose $c \sim p(c)$

choose $z \sim p(z)$

meta-dynamics

action generator
CAVIN: Hierarchical planning in learned latent spaces

Choose $c \sim p(c)$

Choose $z \sim p(z)$

Meta-dynamics

Action generator

Dynamics
Learning with cascaded variational inference

task-agnostic interaction

\[
h(s'' | s, c) = q_h(c | s, s'')
\]

meta-dynamics

\[
g(a | s, c, z) = q_g(z | s, c, a)
\]

action generator
kinect2 sensor

visual observation

preprocess

$S_t$

CAVIN Planner

action $[x, y, \Delta x, \Delta y]$
Tasks

- **clearing**
  
  Clear all objects within the area of blue tiles.

- **insertion**
  
  Move the target to the goal without traversing red tiles.

- **crossing**
  
  Move the target to the goal across grey tiles.
Quantitative Evaluation

Hierarchical Latent space dyn. ↓ Better performance with sparse reward signal

Averaged over 3 Tasks with 1000 test instances each

MPC (Guo et al. ’14, Agrawal et al. ’16, Finn et al. 17); CVAE-MPC (Ichter et al. 18), SeCTAR (Co-Reyes et al ‘18)
Move 2 obstacles

5x
Get around
5x

Open path
Compositional Representations

- Learning Disentanglement
- Learning Keypoints
- Learning Causal Graphs
Composition through Keypoints

Interpretatable


Unsupervised

Composition through Keypoints

Interpretable


Unsupervised

Learning Keypoints From Video

The Pose Encoder is run twice

Sample future frame $T_{\text{temp}}(x)$

Apply color jittering $T_{cj}(x)$

Pose Encoder

Fit Gaussian

$\tilde{\Phi}_{cj}$

$\Phi_{cj}$

MaskNet

$M_{cj}$

$M_{cj} \odot \tilde{x}_{fg} + (1 - M_{cj}) \odot \tilde{x}_{bg}$

Composite

$\Phi_{temp}$

1st Layer Feature maps

Appearance Encoder

Foreground Decoder

$\tilde{x}_{fg}$

$\tilde{x}_{bg}$

$1 - M_{temp}$

BGNet

$\odot=$hadamard

Invert
Learning Keypoints From Video

*The Pose Encoder is run twice*
Learning Keypoints From Video
Learning Keypoints From Video

*The Pose Encoder is run twice.*

\[ \Phi_{pose_{cj}} \]

\[ \Phi_{pose_{temp}} \]

Composite

\[ M_{cj} \odot \tilde{x}_f + (1 - M_{cj}) \odot \tilde{x}_b \]

\[ M_{temp} \]

\[ \tilde{x}_b \]

\[ \tilde{x}_f \]

\[ 1 - M_{temp} \]

BGNet
Learning Keypoints From Video

*The Pose Encoder is run twice*
Learning Keypoints From Video
## Learning Keypoints From Video

<table>
<thead>
<tr>
<th>BBC Pose</th>
<th>Acc.</th>
<th>Human3.6M</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charles et al. [3]</td>
<td>79.9%</td>
<td>Newell et al. [18]</td>
<td>2.16</td>
</tr>
<tr>
<td>Pfister et al. [21]</td>
<td>88.0%</td>
<td>Thewlis et al. [33]</td>
<td>7.51</td>
</tr>
<tr>
<td>unsupervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jakab et al. [8]</td>
<td>68.4%</td>
<td>Zhang et al. [41]</td>
<td>4.91</td>
</tr>
<tr>
<td>Lorenz et al. [14]</td>
<td>74.5%</td>
<td>Lorenz et al. [14]</td>
<td>2.79</td>
</tr>
<tr>
<td>Baseline (temp)</td>
<td>73.3%</td>
<td>Baseline (temp)</td>
<td>3.07</td>
</tr>
<tr>
<td>Baseline (temp, tps)</td>
<td>73.4%</td>
<td>Baseline (temp, tps)</td>
<td>2.86</td>
</tr>
<tr>
<td>Ours</td>
<td>78.8%</td>
<td>Ours</td>
<td>2.73</td>
</tr>
</tbody>
</table>

| MAFL              |       |                            |       |
| unsupervised      |       |                            |       |
| Thewlis et al. [33]| 6.32  | Zhang et al. [41]          | 3.46  |
| Jakab et al. [8]  | 3.19  | Baseline (tps)             | 4.34  |
| Baseline (tps)    |       | Ours (No Mask)             | 2.88  |
| Ours              |       | Ours                       | 2.76  |
Unsupervised Keypoints: Batch RL

Large Set of Task Demonstrations

Policy Learning without Interaction
Unsupervised Keypoints: Batch RL
Unsupervised Keypoints: Batch RL

Unsupervised Representation Learning

Accuracy

BC
BCQ
IRIS
Unsupervised Keypoints: Video Prediction
Unsupervised Keypoints: Video Prediction
Compositional Representations

- Learning Disentanglement
- Learning Keypoints
- Learning Causal Graphs
Learning Causality

• Intuitive Physics

• vs Reinforcement Learning
  • Generalization
  • Goal specification
  • Sample efficiency

• vs Analytical Physics Model
  • Underlying dynamics is uncertain or unknown
  • Simulation is too time consuming
  • Partial observation
Learning Causality

$g_0, g_1$: hidden confounder on the edge
Learning Causality

- Intuitive Physics

- State Representation?
  - Keypoints

- vs. 6 DoF pose

Photo from Jakab et al., Unsupervised Learning of Object Landmarks through Conditional Image Generation

Photo from Tremblay et al., Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects
From left to right:
(1) Input image
(2) Predicted keypoints
(3) Overlay
(4) Heatmap from the keypoints
(5) Reconstructed target image
Learning Causality

$G_0, G_1$: hidden confounder on the edge
Learning Causality

Visual Observations → Unsupervised Keypoint Detection → Keypoints $\{o_t^0, o_t^1, o_t^2\}_{t=1}^T$ → Inference Module → Causal Summary Graph $\{o_t^0, o_t^1\}$ → Dynamics Module → Causal Full Time Graph $\{o_t^0, o_t^1, o_t^2\}_{t=1}^{t+1}$

$g_0, g_1$: hidden confounder on the edge
Learning Causality

Ground truth graph

Predicted graph

Predicted keypoint movements

Ground truth graph

Ground truth keypoint movements
Learning Causality: Extrapolation

(a) Accuracy on edge type: {null edge, spring, rigid}
(b) Correlation on the rest length of the spring relation
(c) Correlation on the length of the rigid relation
(d) Mean squared error on future prediction
Learning Causality: Counterfactual

(a) Intervention on the rest length in spring

(b) Intervention on the length of the rigid relation

(c) Intervention on edge type
Learning Causality

Predicted graph

Predicted keypoint movements

Ground truth keypoint movements
Learning Causality

Predicted graph

Predicted keypoint movements

Ground truth keypoint movements
Learning Causality

Predicted graph

Predicted keypoint movements

Ground truth keypoint movements
Unsupervised Representations towards Counterfactual Predictions

garg@cs.toronto.edu // pair.toronto.edu // Twitter: @Animesh_Garg

Animesh Garg