Data Dreaming for Object Detection:
Learning Object-Centric State Representations for Visual Imitation

Maximilian Sieb
Katerina Fragkiadaki
Computer Vision works well for ImageNet and COCO Categories
Problem: Object categories are different
Proposed method: Data Dreaming for Object-Detection

- On-the-fly data augmentation with synthetically generated data
- Robust per-instance object detectors
- May not work on all scene variations, but does very well on the particular environment and its specific variations
1. Obtaining object masks
   - Background subtraction gives ground truth object masks
   - Requires single un-occluded image of each relevant object

2. Creating synthetic data
   - Massive data augmentation of ground truth
   - Overlay images with random transformations and occlusions to obtain ground truth occluded masks

3. Training detectors
   - Mask R-CNN* as architecture of choice for instance segmentation

---

Proposed Method:
Use bounding box to focus on relevant objects in the scene

a. Spatial features capture relations between different objects (pairwise difference of box $x$, $y$-centroid & depth-value)

b. Visual/Appearance features capture object-specific state
Appearance Feature Encoding

- We use **deep features** extracted from the RGB channels of each bounding box.
- We train these features with **unsupervised losses**, e.g. time-constrastive loss (see below), multi-view invariance loss etc. to enforce focusing on the relevant objects.
Our method:

a. **Spatial features** capture relations between different objects

b. **Visual/Appearance features** capture object-specific state
Discriminative Rewards through Box-Centered Features

- Able to distinguish between correct imitations of target object
- Full-frame encoding does not capture change in scene appropriately
We use a trajectory optimization method (PILQR*)

Using visual features as the state representation $x$, minimize visual dissimilarity between demonstrator and imitator as cost $c$

Given: $\tilde{x}_0, \phi(.)$

For $t = 0, 1, 2, \ldots, T$

• Solve

$$\min_{x, u} \sum_{k=t}^{T} c_k(x_k, u_k), \quad c_k = \alpha \frac{1}{2} \|\phi(x_k) - \phi(x_k^{dem})\|_2^2$$

s.t. $x_{k+1} = f(x_k, u_k), \quad \forall k \in \{t, t+1, \ldots, T-1\}$

$x_t = \tilde{x}_t$

• Execute $u_t$

• Observe resulting state $\tilde{x}_{t+1}$

Model-Based Control for Sample-Efficient Learning

Learned policy

Demonstration
We can leverage on-the-fly training of object detectors to obtain a structured state representation

Benefits:

- **Visual robustness**: Robust towards visual changes in scene and partial occlusions
- **Graph-like state representation**: Establishes interpretable relations between objects
- **Sample-efficient policy learning**: Hard visual attention allows for learning a policy in only a few iterations, requiring only a few real-world trajectory rollouts

Data Dreaming for Object Detection: Learning Object-Centric State Representations for Visual Imitation