Efficient Iterative Amortized Inference for Learning Symmetric and Disentangled Multi-Object Representations

ICML 2021

Patrick Emami
Pan He, Sanjay Ranka, Anand Rangarajan

github.com/pemami4911/EfficientMORL

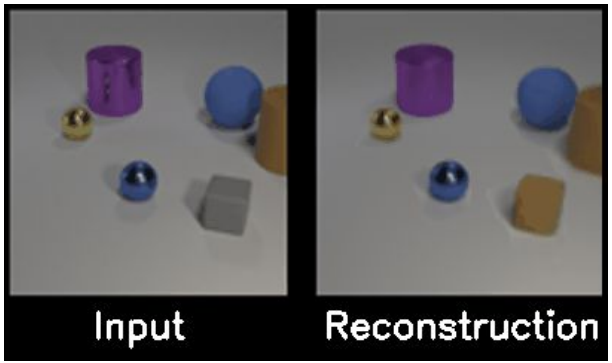
UF UNIVERSITY of FLORIDA

Efficient Multi-**Ob**ject **R**epresentation **L**earning (MORL)

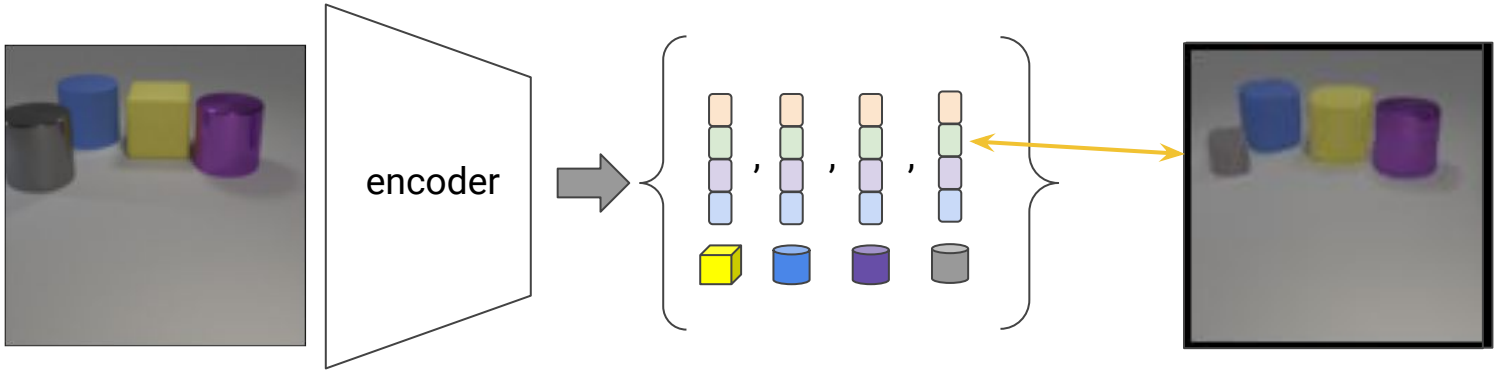
UF UNIVERSITY of FLORIDA

Overview of our paper

- We introduce EfficientMORL, an unsupervised object-centric representation learning framework.
- The learned representations have internal symmetry (common format), are orderless, and disentangle individual object attributes.
- We evaluate the framework on its object decomposition (segmentation), disentanglement, and run time/memory efficiency performance, where it achieves comparative or better results than the relevant prior SOTA method(s) for all metrics.

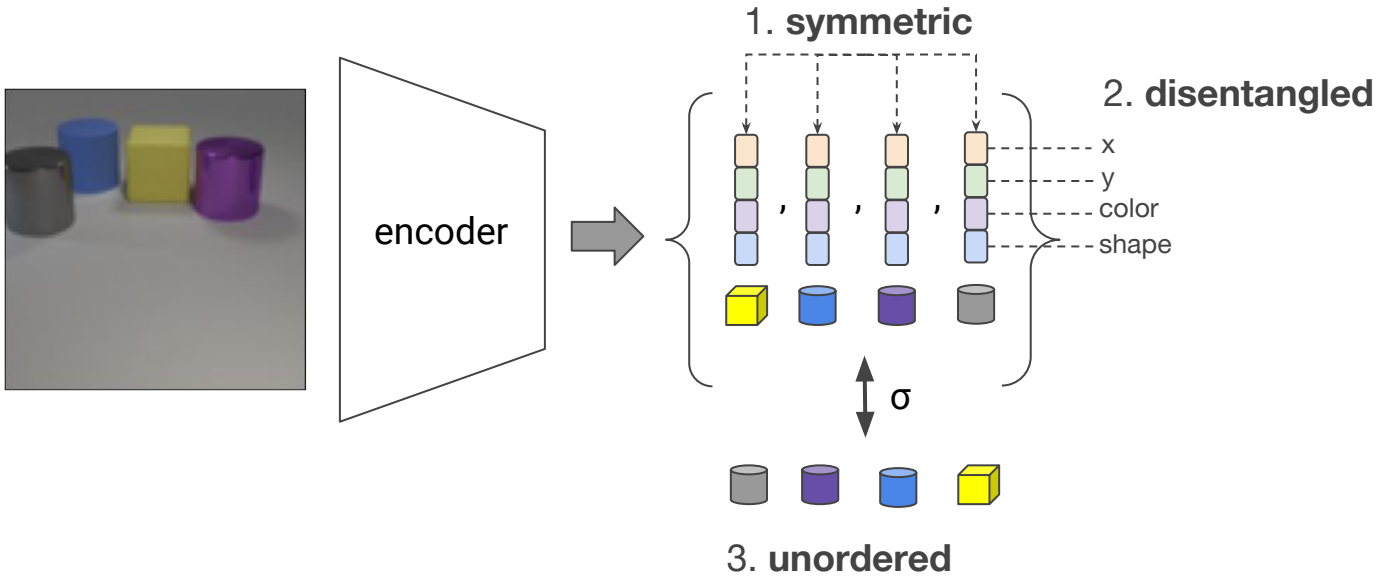


Motivation



Recent object-centric representation learning methods propose to learn set-structured representations of scenes, where each element is an object “slot”.

Motivation



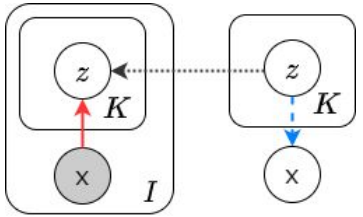
Previous work has established three core properties of object-centric representations. Can we build these three as inductive biases into an **efficient** generative model?

...Yes!

Higgins, Irina, et al. "Towards a definition of disentangled representations." arXiv preprint arXiv:1812.02230 (2018).
Greff, Klaus, Sjoerd van Steenkiste, and Jürgen Schmidhuber. "On the Binding Problem in Artificial Neural Networks." arXiv preprint arXiv:2012.05208 (2020).

A Hierarchical VAE for MORL

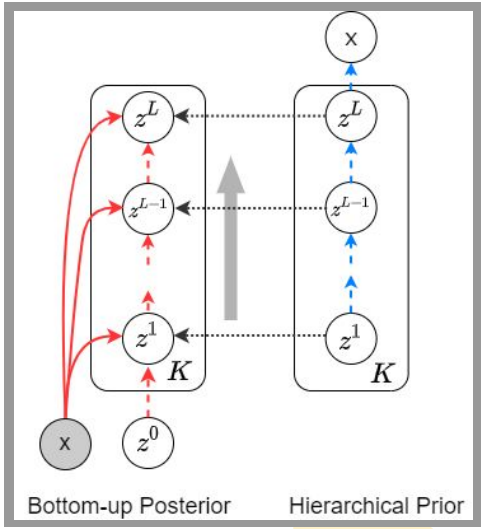
- Images are probabilistic mixtures of *K* RGB components and segmentation masks conditioned on an orderless set of slots $\mathbf{z} := \{z_1, \ldots, z_K\}$.
- Variational posterior for symmetric, disentangled, and orderless slots:
  - $$q_\lambda(\mathbf{z} \mid x) = \prod_{k=1}^{K} q_\lambda(z_k \mid x).$$
- IODINE proposes to use iterative amortized inference to directly refine the Gaussian posterior parameters λ starting from an uninformed guess. Their refinement network takes in **high-dimensional** image-shaped inputs and gradients, and needs **many (5) costly steps** at train and test time.



IODINE (Greff et al., ICML'19)

Key idea #1

Cast the iterative assignment of pixels to slots as bottom-up inference in an HVAE, whose stochastic layers use symmetry-preserving Transformer-like attention to implement the set-to-set mapping.



- Splits the latent variables across *L* layers for *KL* total latents.
- Inference in an HVAE is fast to compute and a hierarchical prior can be used to achieve disentanglement via KL regularization.

Two-stage Inference

- We find that optimizing the HVAE's ELBO directly is challenging, in part because the posterior is highly multimodal due to the orderless property plus the ELBO tries to balance reconstruction quality with hierarchical prior regularization.

Key idea #2

Use only a **few (1-3) steps** of iterative amortized inference to refine the HVAE posterior. The refinement network can then be implemented as a simple recurrent network with **low-dimensional inputs**. The resulting framework thus uses **two-stage inference**.

- That is, for i = 1, …, I,

$$\delta \boldsymbol{\lambda}^{(L,i-1)} = f_\phi(\boldsymbol{\lambda}^{(L,i-1)}, \nabla_{\boldsymbol{\lambda}} \mathcal{L}^{(L,i-1)})$$
$$\boldsymbol{\lambda}^{(L,i)} = \boldsymbol{\lambda}^{(L,i-1)} + \delta \boldsymbol{\lambda}^{(L,i-1)}.$$

- Here, f_ϕ is the recurrent network, with the inputs being Gaussian parameters $\boldsymbol{\lambda}^{(L,i-1)}$ and the gradient of the negative ELBO $\nabla_\lambda \mathcal{L}^{(L,i-1)}$, and which outputs **additive updates**.
- We see a large improvement in stability and convergence speed with just a few (1-3) refinement steps.

Greff, Klaus, et al. "Multi-object representation learning with iterative variational inference." International Conference on Machine Learning. PMLR, 2019.
Marino, Joe, Yisong Yue, and Stephan Mandt. "Iterative amortized inference." In International Conference on Machine Learning, 2018.

Training

- We minimize the following loss:

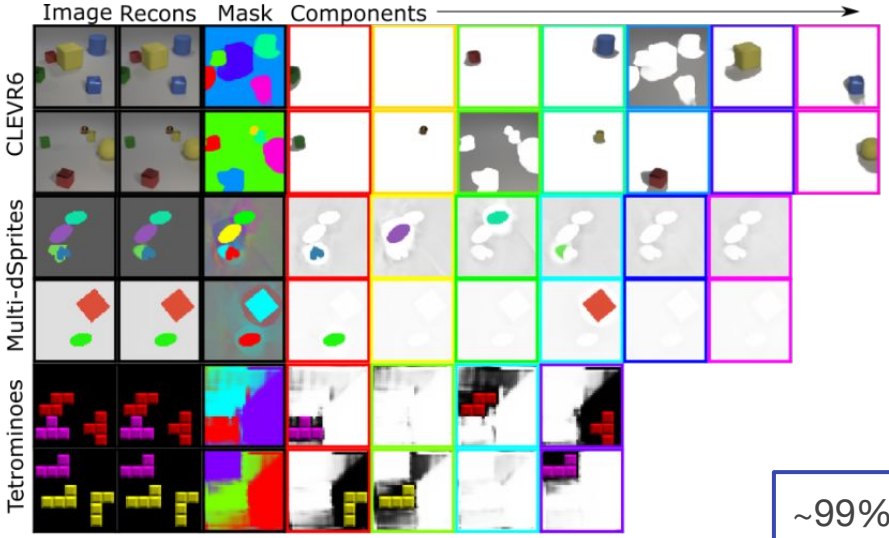
Discounts later refinement steps.

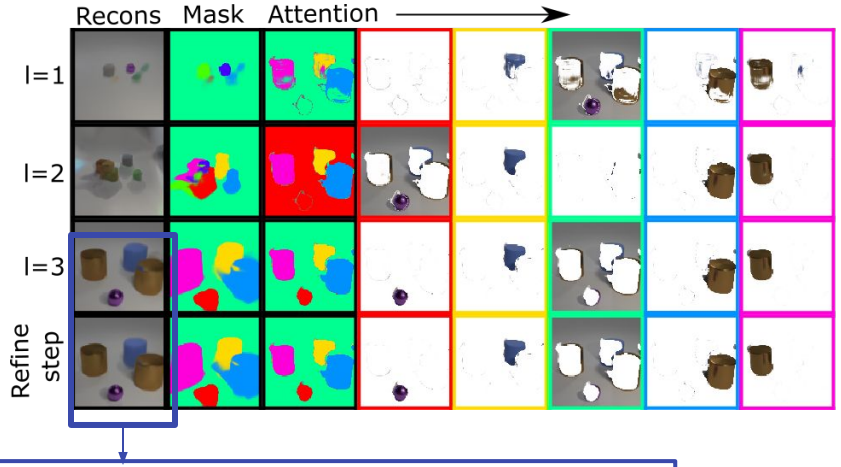
$$\underbrace{\mathcal{L}_{\text{NLL}}^{(L,0)} + D_{\text{KL}}^{(L,0)}}_{\text{Negative ELBO for the HVAE.}} + \sum_{i=1}^{I} \overbrace{\frac{I-(i-1)}{I+1}}^{} \underbrace{\mathcal{L}_{\text{NLL}}^{(L,i)} + D_{\text{KL}}^{(L,i)}}_{\text{Negative ELBO after step } i \text{ of refinement.}}.$$

- We are able to reduce the number of refinement steps *I* after an early phase of training using a **curriculum** to speed up training, and analyze this in depth in the experiments.
  - At test time the refinement stage can optionally be discarded at only small cost in final KL.
- Posterior collapse in the lower layers of the HVAE is mitigated with GECO and by using the KL term in the refinement loss to push lower layers of the posterior to match layer the top layer L.
  - The considered strategies are analyzed via ablation studies.

Rezende, Danilo Jimenez, and Fabio Viola. "Taming vaes." arXiv preprint arXiv:1810.00597 (2018).

Object Decomposition





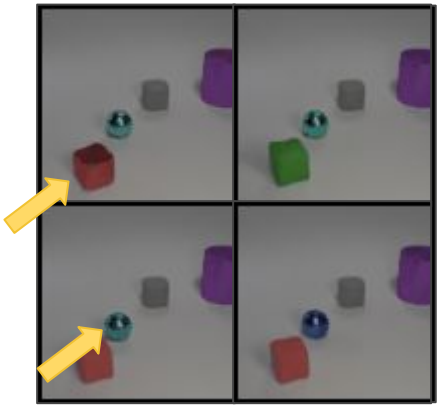
~99% of the refined segmentation and reconstruction achieved with **0** test refinement steps. With 1 step achieves lowest KL.

Multi-object Benchmark. FG-Adjusted Rand Index (ARI) scores (mean ± stddev for five seeds). We replicated Slot Attention's CLEVR6 results over five random seeds (**) but one run failed—without it, the ARI improves from 93.3 to 98.3. Tetrominoes (*) was reported with only 4 seeds.

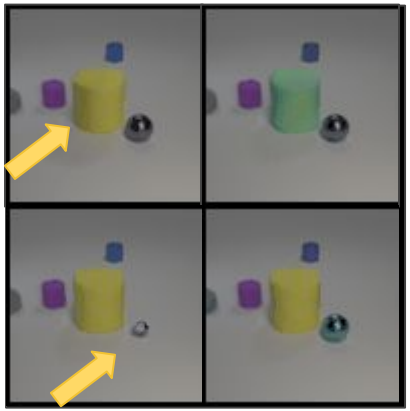
	CLEVR6	Multi-dSprites	Tetrominoes
Slot Attention	98.8 ± 0.3	$\mathbf{91.3 \pm 0.3}$	$99.5 \pm 0.2^*$
Slot Attention (**)	93.3 ± 11.1	—	—
IODINE	$\mathbf{98.8 \pm 0.0}$	76.7 ± 5.6	$\mathbf{99.2 \pm 0.4}$
MONet	96.2 ± 0.6	90.4 ± 0.8	—
Slot MLP	60.4 ± 6.6	60.3 ± 1.8	25.1 ± 34.3
EfficientMORL	96.2 ± 1.6	$\mathbf{91.2 \pm 0.4}$	98.2 ± 1.8

Disentanglement

- Slot Attention is efficient and symmetric, yet is a deterministic autoencoder.
- EfficientMORL's representations have fewer correlated, redundant, and active latent dimensions.



Slot Attention



EfficientMORL

DCI on CLEVR6 (mean ± std dev across 5 runs). Higher is better.

	Disentanglement	Completeness	Informativeness
Slot Attention	0.46 ± 0.01	0.38 ± 0.02	0.34 ± 0.01
EfficientMORL	$\mathbf{0.63} \pm 0.04$	$\mathbf{0.63} \pm 0.06$	$\mathbf{0.46} \pm 0.01$

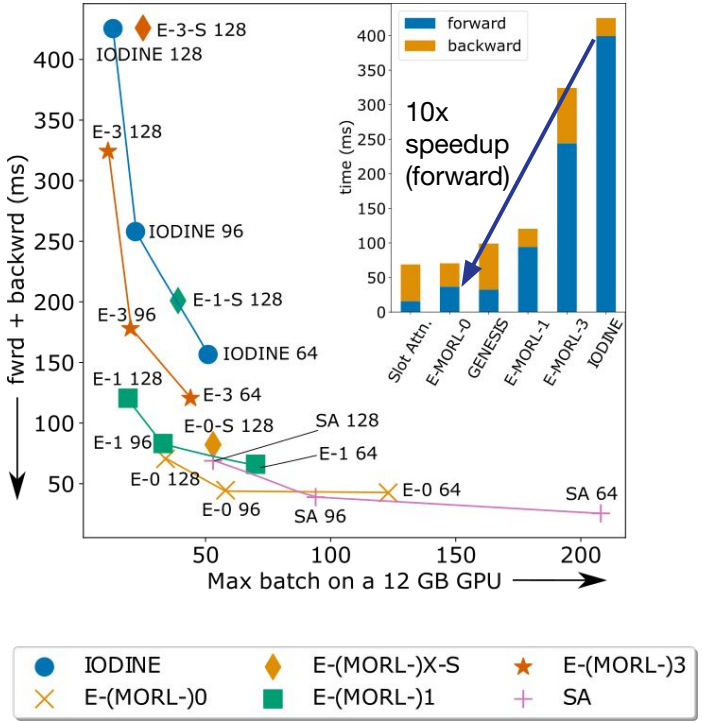
Locatello, Francesco, et al. "Object-centric learning with slot attention." arXiv preprint arXiv:2006.15055 (2020).

Run-time and Memory Benchmarking

Comparisons on forward and backward pass timing and memory consumption provided in paper.

Wall clock training time

- Using a curriculum on refinement steps during training, CLEVR6 with 96 x 96 images takes **~17 hours.**
- Fully training IODINE would take **~1 week** using on the same hardware, almost 10x slower.



Lines connect results for square images with side length 64, 96, and 128 for each model. "E-X-S" is ours with Slot Attention's memory-efficient decoder.

Conclusions

- We introduced EfficientMORL, a multi-object scene representation learning framework.
- We hope that in the short term, our work will expedite research on this exciting topic due to faster training/test times, improved stability, and lower memory consumption---particularly for more compute intensive problems, e.g., videos.

Github