SPOT: Self-Training with Patch-Order Permutation for Object-Centric Learning with Autoregressive Transformers

Ioannis Kakogeorgiou[1] Spyros Gidaris[2] Konstantinos Karantzalos[1] Nikos Komodakis[3,4,5]

[1]National Technical University of Athens [2]valeo.ai
[3]University of Crete [4]IACM-Forth [5]Archimedes/Athena RC

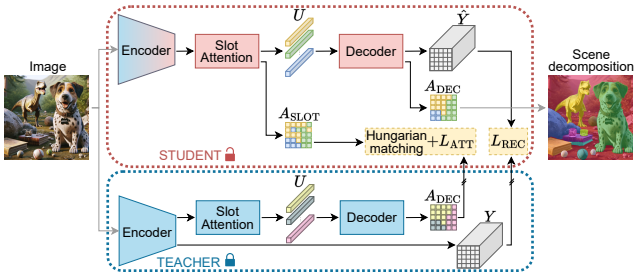
Abstract

Unsupervised object-centric learning aims to decompose scenes into interpretable object entities, termed slots. Slot-based auto-encoders stand out as a prominent method for this task. Within them, crucial aspects include guiding the encoder to generate object-specific slots and ensuring the decoder utilizes them during reconstruction. This work introduces two novel techniques, (i) an attention-based self-training approach, which distills superior slot-based attention masks from the decoder to the encoder, enhancing object segmentation, and (ii) an innovative patch-order permutation strategy for autoregressive transformers that strengthens the role of slot vectors in reconstruction. The effectiveness of these strategies is showcased experimentally. The combined approach significantly surpasses prior slot-based autoencoder methods in unsupervised object segmentation, especially with complex real-world images. We provide the implementation code at https://github.com/gkakogeorgiou/spot.

Slot-based auto-encoding frameworks stand at the forefront of object-centric learning approaches. These frameworks typically rely on two main components. The first main component is a slot-attention encoder that generates (though an attention-like mechanism) a set of latent vectors referred to as 'slots', each intended to represent an individual object within an image. The second main component is an autoregressive decoder burdened with the challenging task of reconstructing the input based on information derived from the extracted slots, guiding the learning of object-centric representations. SPOT provides significant improvements with respect to both of these components.

Improving slot generation through self-training: slot-based attention masks can be generated by both encoder and decoder. Building on the insight that masks produced during decoding demonstrate superior object decomposition, SPOT proposes a self-training scheme that distills slot-based attention masks from the decoder to the encoder, enhancing object segmentation information captured by slots.

Enhanced autoregressive decoders with sequence permutations: in autoregressive transformers, initial patch tokens rely heavily on slot vectors due to limited context. As decoding progresses, earlier token dependence grows, weakening slot encoder learning. SPOT introduces sequence permutations, which helps amplifying the role of slot vectors in the reconstruction process and leads to a more robust supervisory signal for object-centric learning.



SPOT's self-training approach.

The synergistic application of these strategies enables SPOT to achieve state-of-the-art results in real-world object-centric learning. On COCO, SPOT surpasses the prior state-of-the-art by 2.7 and 5.9 points in MBOi and MBOc metrics, showcasing its superiority.

METHOD	COCO		PASCAL	
	MBO^i	MBO^c	MBO^i	MBO^c
SlotAttention	17.2	19.2	24.6	24.9
SLATE	29.1	33.6	35.9	41.5
CAE	-	-	32.9$_{\pm 0.9}$	37.4$_{\pm 1.0}$
DINOSAUR	32.3$_{\pm 0.4}$	38.8$_{\pm 0.4}$	44.0$_{\pm 1.9}$	51.2$_{\pm 1.9}$
DINOSAUR-MLP	27.7$_{\pm 0.2}$	30.9$_{\pm 0.2}$	39.5$_{\pm 0.1}$	40.9$_{\pm 0.1}$
Rotating Features	-	-	40.7$_{\pm 0.1}$	46.0$_{\pm 0.1}$
SlotDiffusion	31.0	35.0	**50.4**	55.3
(Stable-)LSD	30.4	-	-	-
SPOT w/o ENS (ours)	34.7$_{\pm 0.1}$	44.3$_{\pm 0.3}$	48.1$_{\pm 0.4}$	55.3$_{\pm 0.4}$
SPOT w/ ENS (ours)	**35.0**$_{\pm 0.1}$	**44.7**$_{\pm 0.3}$	48.3$_{\pm 0.4}$	**55.6**$_{\pm 0.4}$

Comparison with object-centric methods on COCO and PASCAL.

SPOT enhances unsupervised object-centric learning in slot-based autoencoders, excelling in complex real-world images.

Relation to the workshop: SPOT advances unsupervised object-centric learning with attention-based self-training & patch-order permutation, achieving state-of-the-art results.
Publication: Our work has been selected as a highlight and will be published in the proceedings of CVPR 2024.