SORNet: Spatial Object-Centric Representations for Sequential Manipulation

Wentao Yuan[1] Chris Paxton[2] Karthik Desingh[1] Dieter Fox[1,2]

[1]University of Washington [2]NVIDIA

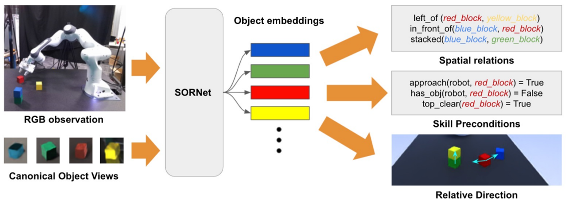
PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

THE UNIVERSITY OF UTAH

NVIDIA

Overview

We propose **SORNet**: **S**patial **O**bject-centric **R**epresentation **Net**work to learn object-centric embeddings that encode spatial relationships

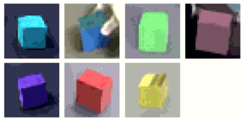

RGB observation
Canonical Object Views
SORNet
Object embeddings
left_of (red_block, yellow_block)
in_front_of(blue_block, red_block)
stacked(blue_block, green_block)
Spatial relations
approach(robot, red_block) = True
has_obj(robot, red_block) = False
top_clear(red_block) = True
Skill Preconditions
Relative Direction

Key Features

SORNet generalizes **zero-shot** to scenes with unseen objects and different number of objects.

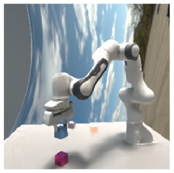
Training objects

Testing objects



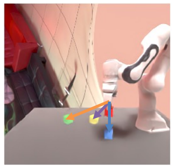
SORNet is trained only on classification of **logical** predicates but captures **continuous** spatial relationships.

Training objective (logical)



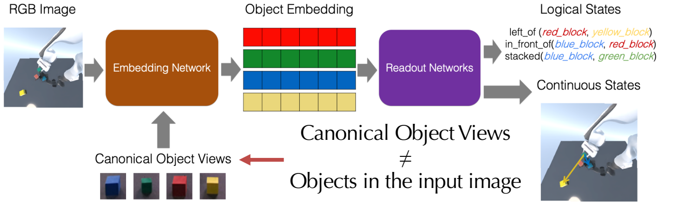
aligned_with(cobalt_block, ruby_block)
has_obj(robot, cobalt_block)
in_approach_region(robot, ruby_block)
on_surface(lavender_block, right)
on_surface(peach_block, right)
on_surface(ruby_block, left)
top_is_clear(cobalt_block)
top_is_clear(lavender_block)
top_is_clear(peach_block)
top_is_clear(ruby_block)

Testing objective (continuous)



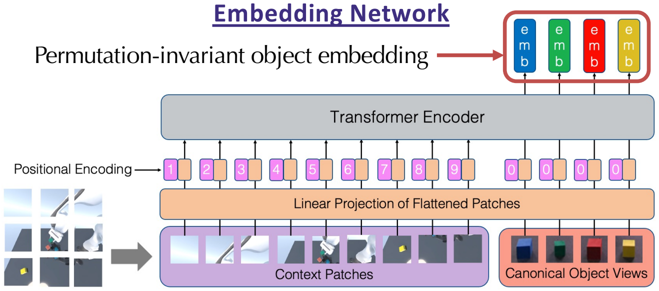
Method

SORNet: System Overview



Canonical Object Views ≠ Objects in the input image

Embedding Network

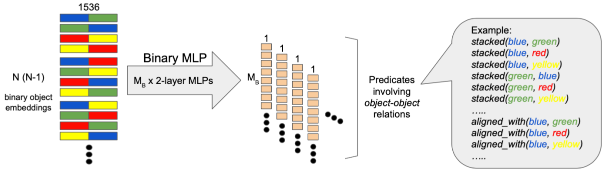


Permutation-invariant object embedding

Learned Attention

Readout Networks

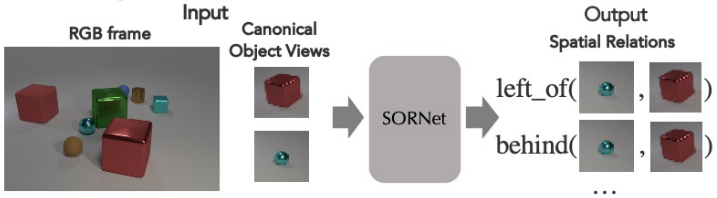
Number of outputs changes adaptively with number of inputs



Example:
stacked(blue, green)
stacked(blue, red)
stacked(blue, yellow)
stacked(green, blue)
stacked(green, red)
stacked(green, yellow)
…..
aligned_with(blue, green)
aligned_with(blue, red)
aligned_with(blue, yellow)
…..

EXPERIMENTS

Spatial Relation Prediction on CLEVR-CoGenT



Training (Condition A)

- Cubes are gray, **blue**, **brown**, or **yellow**
- Cylinders are **red**, **green**, **purple**, or **cyan**
- Spheres can have any color

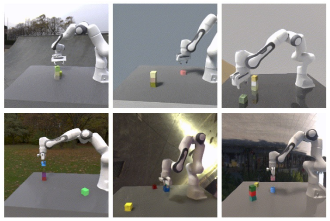
Testing (Condition B)

- Cubes are **red**, **green**, **purple**, or **cyan**
- Cylinders are gray, **blue**, **brown**, or **yellow**
- Spheres can have any color

Zero-shot Accuracy

	MDETR [34]	MDETR-oracle [34]	**SORNet(ours)**
ValA Accuracy	84.950	97.944	**99.006**
ValB Accuracy	59.627	98.052	**98.222**

Predicate Classification on Leonardo



Training split

- ➢ Overall 405 colored blocks.
- ➢ Randomly chosen 4 blocks in each sequence
- ➢ One task - stacking 4 blocks.
- ➢ 133796 sequences.

Testing split

- ➢ 7 colored blocks (unseen in train)
- ➢ Randomly chosen 4-6 blocks in each sequence
- ➢ 7 tasks different from training
- ➢ 9526 sequences.

Objective

- ➢ Classifying logical predicates from RGB input

Method	# pred	F-1 Score all	on_surface	has_obj	top_clear	stacked	aligned	approach
ResNet18 M-Head 100-shot	52	0.0	21.9	0.0	32.6	0.0	0.0	0.0
ViT-B/32 M-View 100-shot	52	0.0	37.7	6.3	46.5	0.0	0.0	7.3
ViT-B/32 M-Head M-View 100-shot	52	0.0	70.5	31.0	73.2	27.2	0.0	23.2
SORNet 0-shot	52	83.2	92.2	79.7	93.0	91.2	63.8	74.9
SORNet M-View 0-shot	52	88.9	**97.5**	82.0	**98.4**	**97.3**	**70.5**	**81.7**
SORNet M-View (G) 0-shot	52	**89.5**	97.1	**94.7**	96.8	96.4	69.9	76.7
SORNet M-View (G) 5 obj 0-shot	70	85.3	96.0	96.7	91.3	83.6	69.8	78.1
SORNet M-View (G) 6 obj 0-shot	102	79.9	95.5	97.0	87.5	69.2	70.0	77.9

2022 NRI & FRR Principal Investigators' Meeting
April, 2022

Award ID#: 2024778