Learning the consequences of actions

ACT-Thor: A Controlled Benchmark for Embodied Action Understanding in Simulated Environments

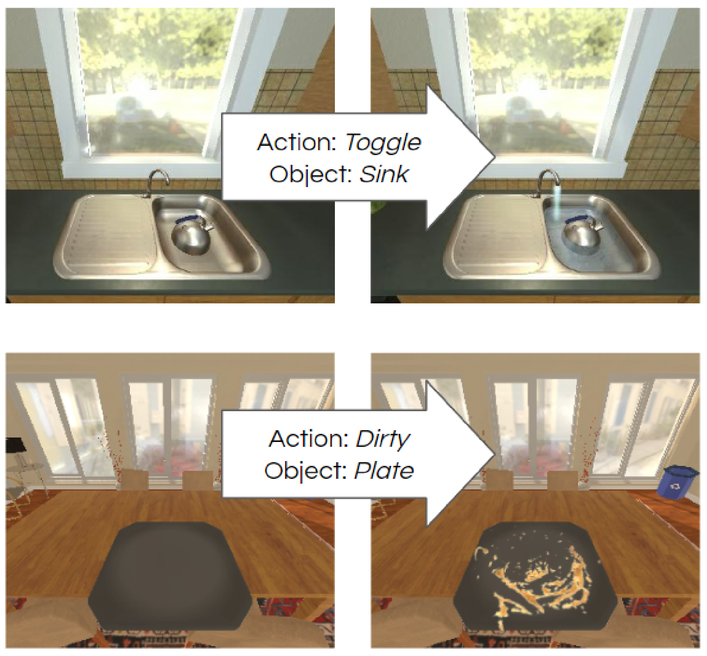
Michael Hanna, Federico Pedeni, Alessandro Suglia, Alberto Testoni, Raffaella Bernardi

m.w.hanna@uva.nl, federico.pedeni@studenti.unitn.it, a.suglia@hw.ac.uk, alberto.testoni@unitn.it, raffaella.bernardi@unitn.it

University of Amsterdam; University of Trento; Heriot-Watt University

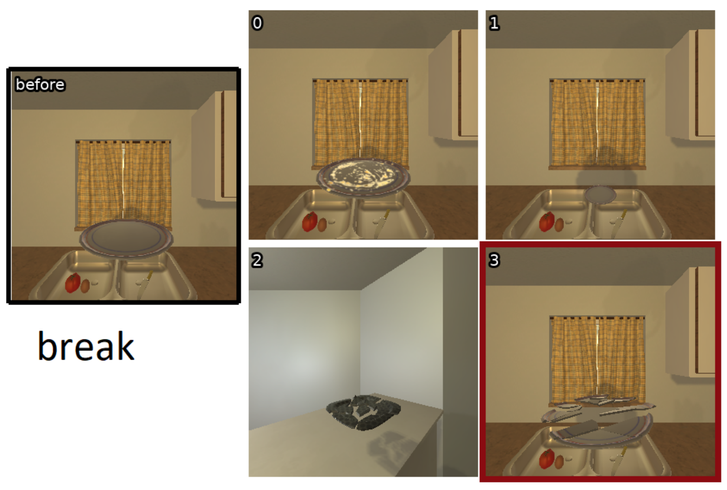
Action Understanding

Embodied agents should understand their actions' effects on their environment. We test model understanding via a new dataset!



AI2-Thor [1] is a platform where an agent (robot) interacts with a virtual environment. We use it to generate images for our dataset:

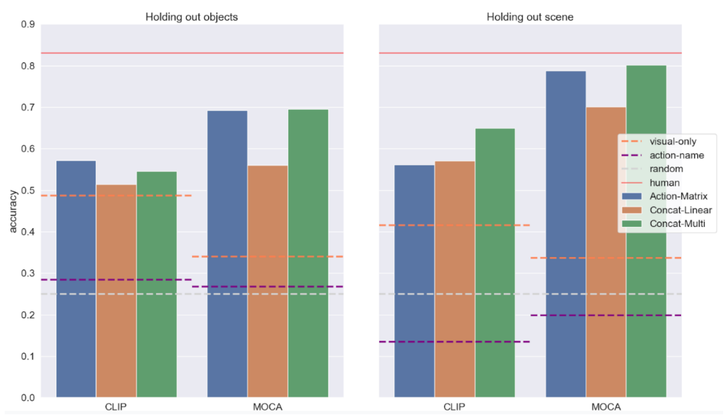
- Select an object; either pick up or place it.
- Record the before image of the object.
- Perform an action on the object
- Record the after image of the object.



We arrange images into **contrast sets**. Given the "before-image" and action label, the model must choose which "after-image" is the true result from performing the given action in the given scene.

Evaluation

We use this dataset to evaluate SotA visual encoders, as part of simple baselines that take in the action and a representation of the before-image, and predict a representation of the after-image



We test models on predicting the outcome of actions on unknown objects or in unknown scenes. Some generalize well to new scenes, but none to new objects.

Action	Nearest neighbors (sorted)		
break	dirty	open	toggle
close	break	dirty	put
dirty	break	pull	open
drop	push	pull	pickUp
fill	put	pull	throw
open	dirty	break	fill
pickUp	dirty	fill	break
pull	put	push	throw
push	pull	throw	put
put	throw	pull	push
throw	put	push	pull
toggle	dirty	break	pull

We examine the action representations learned by our models for semantic clusters

Takeaways

- We can use virtual environments as controlled settings for dataset generation!
- We create a dataset for learning actions and consequences with AI2-Thor. Our task is easy for humans, with high agreement.
- However, the task is not as easy for simple baselines, especially with new objects.

Dataset Details

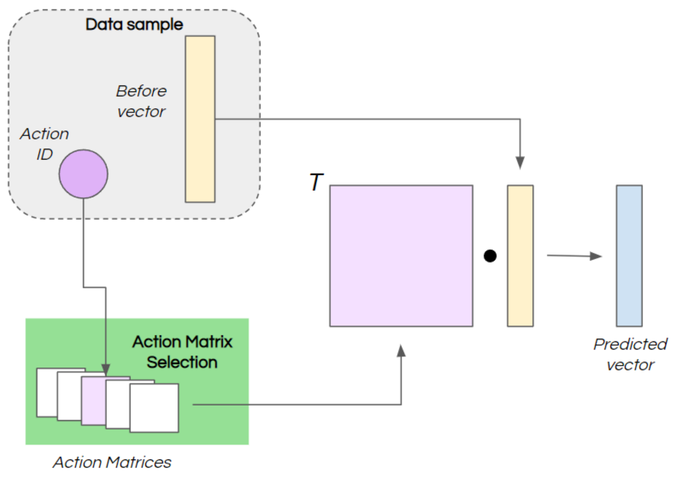
We amass a dataset of contrast sets. Though small, this can be expanded, and the data repurposed for other tasks.

Statistic	Count
action-object pairs	403
before-i, action, after-i	11154
unique before-i	3746
unique after-i	11154
scenes	120
objects	62
actions	12

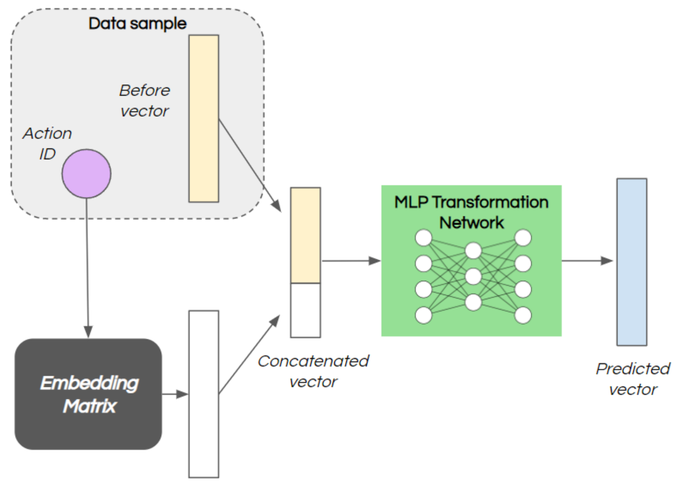
We solicit human annotations: accuracy is 81%, and agreement is 85%.

Model Details

Models extract visual features from the before-image via MOCA [2] or CLIP [3]. The probability that a given after-image is correct is modeled as the similarity between the predicted representation and after-image representation.



Our first model, based on Baroni and Zamparelli (2010) [4], envisions actions as a matrix that transform a scene.



Our second and third models use a linear layer or MLP to transform a joint image-action embedding into a prediction vector.

[1] Eric Kolve et al. 2017. AI2-Thor: An interactive 3d environment for visual ai. ArXiv, abs/1712.05474.
[2] Kunal Pratap Singhet al. 2021. Factorizing perception and policy for interactive instruction following. In Proceedings of ICCV, pages 1888–1897.
[3] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In Proceedings of the 38th ICML, volume 139 of PMLR, pages 8748–8763. PMLR.
[4] Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In Proceedings of EMNLP 2010, pages 1183–1193.
Poster template by Mike Morrison, Rafael Bailo, Tom Kocmi.