Compositional Learning for Vision-Language Reinforcement Learning Agents

Zijun Lin[1], Haidi Azaman[2], M Ganesh Kumar[3,4], Cheston Tan[4]
Nanyang Technological University[1], National University of Singapore[2], Harvard University[3], Centre for Frontier AI Research, A*STAR[4]

CVPR JUNE 17-21, 2024

Scan me to access pre-print

Summary

Motivation: Children can rapidly generalize compositionally-constructed rules to unseen test sets. On the other hand, deep reinforcement learning (RL) agents need to be trained over millions of episodes, and their ability to generalize to unseen combinations remains unclear.

Problem: We investigate the compositional generalization capabilities for vision-language multimodal learning, using the task of navigating to instructed targets in synthetic 3D environments. Specifically, the instruction contains five colors (**C**) -- red, green, blue, yellow, black and five shapes (**S**) – capsule, cube, cylinder, prism, sphere.

Contribution: We are the first to demonstrate that RL agents can be trained to implicitly learn concepts and compositionality, to solve more complex 3D environments in zero-shot without needing additional training episodes.

Task: Given an instruction with vision inputs, the RL agents are expected to navigate to the target objects.

Performance Criterion: The agent achieves +9 average reward (maximum +10) over 100 episodes, the details of the reward system in our environments are shown in Table 2.

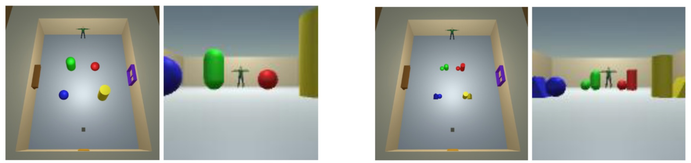


Figure 1. Example Environment.

Environments for Grounded Learning

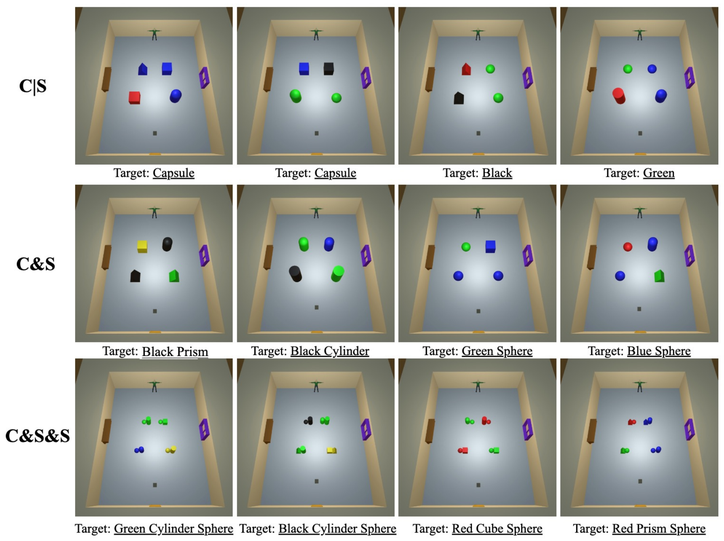
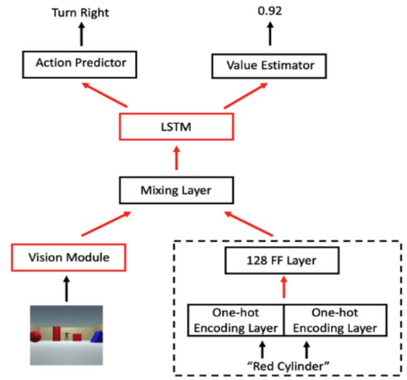


Figure 2. Examples of instruction and various scenarios in three environments.

Experiment I: Generalization of Compositional Learning



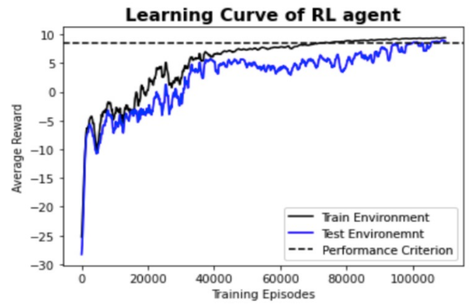


Figure 3. Left: Agent architecture. The language module of the one-hot encoder agent is at bottom right, boxed up within dashed lines. Red arrows or boxes represent trainable weights, while black arrows or boxes represent frozen weights. Right: Learning curve of the RL agent in train environment and test environment respectively.

Shape\Color	Red	Green	Blue	Yellow	Black
Capsule	**Test**	Train	Train	Train	Train
Cube	Train	**Test**	Train	Train	Train
Cylinder	Train	Train	**Test**	Train	Train
Prism	Train	Train	Train	**Test**	Train
Sphere	Train	Train	Train	Train	**Test**

Table 1. Train-Test split for environment C|S and C&S.

Actions	Rewards
Hit the wall	-1
Choose wrong objects	-3
Reach maximum 500 steps	-10
Choose the correct object	+10

Table 2. Reward System.

Experiment II: Concept Learning Speeds Up and Generalizes Compositional Learning

	Episodes (K) for performance criterion	
Training Environment	**Train combinations**	**Held-out Test combinations**
C&S	67.4 ± 7.2	94.8 ± 3.7
C|S → C&S	0.6 ± 0.1	5.5 ± 2.9

Table 3. Comparisons of training episodes needed for agents to reach performance criterion in C&S environments.

		Average reward for zero-shot Evaluation in Environments			
Training Environment	**Training Episodes (K)**	**Familiar C|S combo**	**Unseen C|S combo**	**Familiar C&S&S combo**	**Unseen C&S&S combo**
Nil	0.0	-24.42 ± 1.29	-23.42 ± 2.57	-29.15 ± 3.07	-36.48 ± 3.60
C&S	67.4	0.37 ± 1.50	3.08 ± 0.32	2.84 ± 0.92	-5.10 ± 2.59
C&S	168.6	-8.02 ± 3.01	-2.05 ± 1.57	-8.27 ± 3.75	-23.2 ± 9.97
C|S	168			1.19 ± 1.24	-4.02 ± 2.44
C|S → C&S	168 → 0.6	$\mathbf{8.74 \pm 0.29}$	$\mathbf{7.58 \pm 0.32}$	$\mathbf{5.49 \pm 0.26}$	$\mathbf{5.55 \pm 0.39}$

Table 4. Summary of zero-shot evaluation experiments.

Experiment III: Comparison of Text Encoders

	Training Episodes (K)	
Text Encoder	**Train combinations**	**Test combinations**
One-hot	67.4 ± 7.2	94.8 ± 3.7
Vanilla	116.2 ± 15.4	185.9 ± 15.5
BERT	109.0 ± 9.1	≥ 200
CLIP	$\mathbf{56.2 \pm 5.3}$	$\mathbf{72.6 \pm 6.0}$

Table 5. Comparisons of training episodes needed for agents with different text encoders in C&S environment.

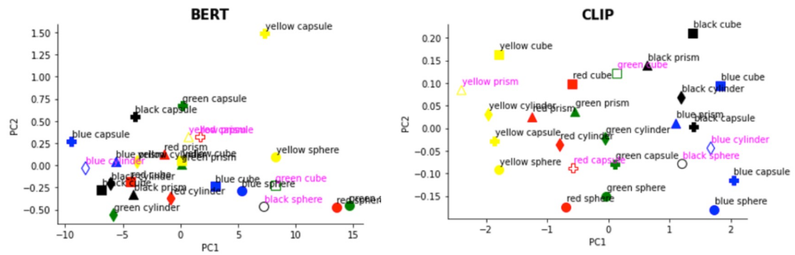


Figure 4. Word embeddings of the agent with BERT (left) and CLIP (right) text encoders after training in the C&S environment for 50,000 episodes. Filled icons represent training set examples, while unfilled icons with magenta labels represent testing set examples.

Conclusions

1. We demonstrated the compositional abilities of reinforcement learning agents with multimodality. Specifically, we found that agents can learn to decompose and recompose instructions to solve held-out test instructions.
2. We showed that invariant concept learning accelerates compositional learning.
3. We tested various text encoders, with CLIP as a foundation model on both image and text modality showing the ability to speed up learning.

Acknowledgements

This work was supported by A*STAR through a CRF award (C.T.), as well as ARIA (H.A.) and CIARE (Z.L.) internships

References

[1] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: Abrief survey. IEEE Signal Processing Magazine, 34(6):26–38, 2017.
[2] Andrew G Barto and Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning. Discrete Event Dynamic Systems, 13(1-2):41–77, 2003.
[3] Felix Hill, Stephen Clark, Phil Blunsom, and Karl Moritz Hermann. Simulating early word learning in situated connectionist agents. In Annual Meeting of the Cognitive Science Society, 2020.
[4] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rockt¨aschel. A survey of zero-shot generalisation in deep reinforcement learning. Journal of Artificial Intelligence Research, 76:201–264, 2023.
[5] James L McClelland. Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. Journal of Experimental Psychology: General, 142(4):1190, 2013.