Toward Compositional Generalization in Object-oriented World Modeling

Linfeng Zhao [1] Lingzhi Kong [1] Robin Walters [1] Lawson L.S. Wong [1]

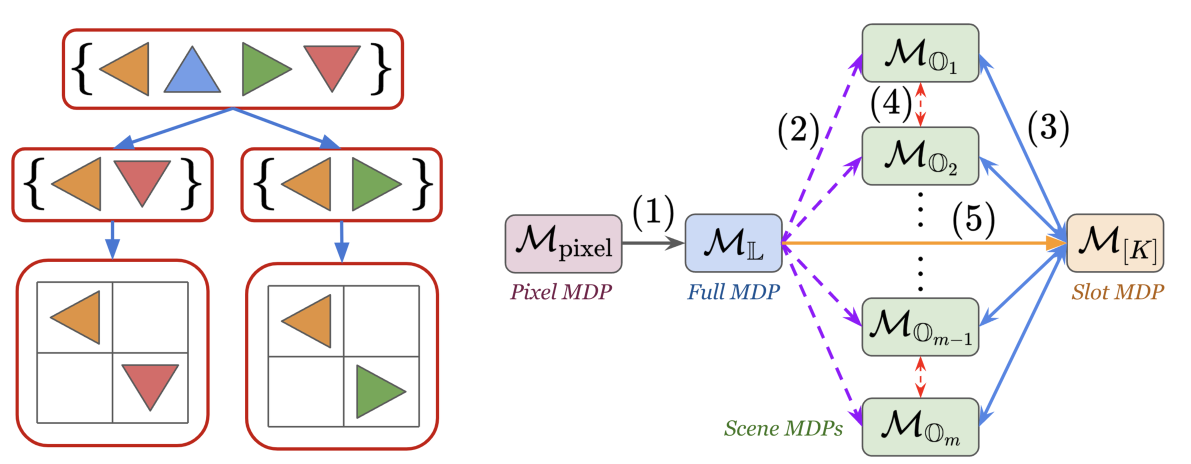
[1]Khoury College of Computer Sciences, Northeastern University

Overview

- We focus on the setting of world modeling in *object-oriented environments* to study *compositional generalization*.
- We (1) formalize the compositional generalization problem with an *algebraic* approach and (2) study how a world model can achieve that.
- We introduce a conceptual environment, Object Library, and two instances, and deploy a principled pipeline to measure the generalization ability.
- Motivated by the formulation, we analyze several methods with *exact* or *no* compositional generalization ability using our framework.
- We design a differentiable approach, Homomorphic Object-oriented World Model (HOWM), that achieves approximate but more efficient compositional generalization.

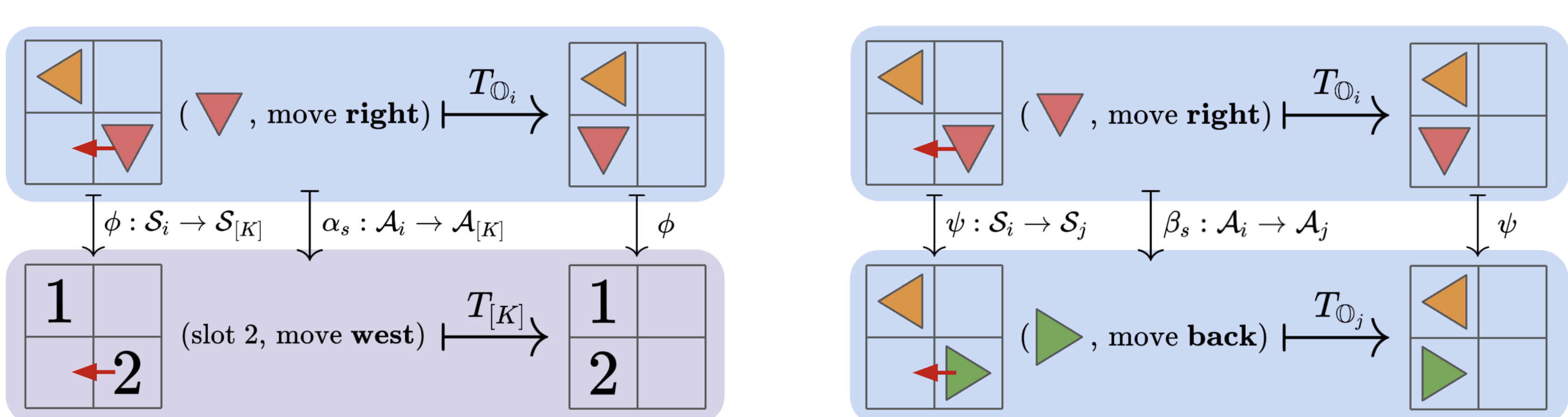
Setup: Object-oriented Environments

Object-oriented Environment: Object Library



- Object library is a conceptual environment, equipped with a "vocabulary" of N objects $\mathbb{L}$, such as {▲, ▼, ◀, ▶}.
- A combination of K objects is a scene (similar to words forming sentences [1]) and forms a separate *scene* MDP.
- All combinations: { {▲, ▼}, {▲, ◀}, {▲, ▶}, {▼, ◀}, {▼, ▶}, {◀, ▶} }.

Compositional Generalization in World Modeling



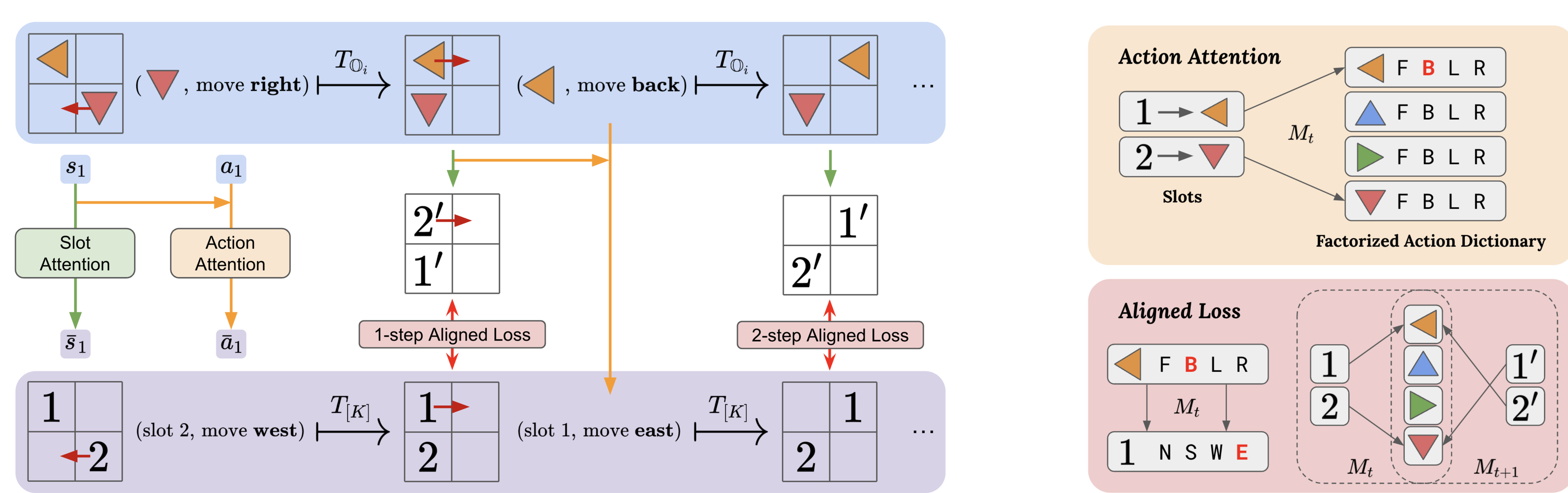
- Setup: *end-to-end* learn a (deterministic) world model $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ in environments with multiple objects (or object-oriented environments).
- Goal: the model T has the ability of *compositional generalization*.
- Challenge: *end-to-end* solve *binding* of N objects and their actions correctly.

Results

CG Type	Env=Shapes	**Eval MRR** (%, 1-step)				**Eval MRR** (%, 5-step)				**Train MRR** (%, 5-step)				**Gap** (MRR %, 5-step)				Memory
(1. Exact CG)	Σ_N-CSWM	100.	100.	99.9	*OM*	99.9	99.9	99.9	*OM*	100.	100.	100.	*OM*	0.0	0.0	0.1	*OM*	8.1GB
(2. No guaranteed CG)	Σ_K-CSWM	100.	56.4	70.3	94.5	99.2	17.9	27.0	64.8	100.	100.	100.	100.	0.8	82.1	73.0	35.2	1.5GB
	Σ_K-CSWM(CA)	97.3	80.0	81.2	76.2	87.7	42.9	43.6	36.1	98.2	99.1	97.0	94.2	10.5	56.2	53.3	58.1	1.6GB
	C-WM(N)	54.3	81.1	65.1	20.2	24.3	72.0	52.1	11.0	72.7	92.5	73.1	42.8	48.4	20.5	20.9	31.8	1.3GB
	MONet(N)+BM	12.6	73.9	35.9	*OM*	2.0	20.2	55.9	*OM*	7.0	64.5	84.8	*OM*	5.0	44.3	29.0	*OM*	9.3GB
(3. Approximate CG)	**HOWM (ours)**	99.2	98.5	99.7	99.7	92.3	84.2	75.1	81.8	97.0	96.9	98.0	98.1	4.7	12.7	22.9	16.3	3.7GB

Method: Compositionally Generalizable World Model

- Our framework draw two possible paths for compositional generalization in world modeling: *exact* and *approximate*. The *exact* approach requires Σ_N-equivariance.
- It is computationally expensive, while we propose a method to provably provide compositional generalization with Σ_K-equivariance, using end-to-end learned binding from interaction data (s, a, s').
- It comes from a corollary of the right proposition, which measures compositional generalization of with equivariance error and related the errors between the model on all objects (N) or scene objects (K).



- **(Left)** Upper blue sequence: a ground pixel MDP with some objects $\mathbb{O}_i$. Lower purple sequence: the slot MDP.
- *Two facts*: (1) encoded object slots in different steps may have different ordering (marked as $1'$ and $2'$), and (2) the transition model is equivariant in slot ordering, i.e., consistent across time steps (in 1 and 2), thus the loss computation needs alignment of slots (between $1, 2$ and $1', 2'$).
- **(Top right)** Action Attention learns to bind actions from interaction (action-object correspondence is *unknown* and learned).
- **(Bottom right)** In the Aligned Loss, the learned binding matrices M_t and M_{t+1} are used to lift slots in t and $t+1$ to a canonical space (full MDP). Positive term (as [2]):

$$\mathcal{L}^+(s_t^\uparrow, s_{t+1}^\uparrow) = \|\mathtt{NG}(M_{t+1}^+)\bar{s}_{t+1} - \mathtt{NG}(M_t^+)T(\bar{s}_t, \bar{a}_t)\|^2 , \qquad (1)$$

Compositional Generalization through Σ_K-equivariance

$\hat{T}_{[K]}$, the induced transition model of $\mathcal{M}_{[K]}$ under $h = \langle \phi, \{\alpha_s \mid s \in \mathcal{S}\}\rangle$, has *sample* equivariance error at $(\phi(s), \alpha_s(a), \phi(s')) \in \mathcal{S}_{[K]} \times \mathcal{A}_{[K]} \times \mathcal{S}_{[K]}$ and $\bar{\sigma} \in \Sigma_K$:

$$\lambda^{\bar{\sigma}}_{[K]} \triangleq \left[\left|\hat{T}_{[K]}(\phi(s') \mid \phi(s), \alpha_s(a))\hat{T}_{[K]}(\bar{\sigma}.\phi(s') \mid \bar{\sigma}.\phi(s), \bar{\sigma}.\alpha_s(a))\right|\right] = C \cdot \lambda^{\sigma}_{\mathbb{L}} , \qquad (2)$$

where $C = \binom{N}{K}$ is the number of K-slot scenes given an N-object library, $\phi : \mathcal{S}_{\mathbb{L}} \to \mathcal{S}_{[K]}$ and $\alpha_s : \mathcal{A}_{\mathbb{L}} \to \mathcal{A}_{[K]}$. The equivariance error is then $\lambda_{[K]} = \mathbb{E}_{s,a,s',\bar{\sigma}}[\lambda^{\bar{\sigma}}_{[K]}] = C \cdot \lambda_{\mathbb{L}}$.

Empirical Analysis

Experimental Setup

- *Shapes* is an instance of the Object Library, built upon the 2-D shape version of the Block Pushing environment [2].
- Compared 3 classes of methods in terms of compositional generalization (CG) ability: (1) *exact*, (2) *no guaranteed*, (3) *approximate* (ours).

Results and Analysis

- Results for all methods on the Shapes environment with $K = 5$ and $N = 5, 10, 20, 30$ (four numbers in each cell). "OM" stands for out of GPU memory, where we limit the usage to 10GB. We report for memory usage on $N = 20$.
- The *exact* approach, Σ_N-CSWM (using N-slot of [2]) can perform near perfectly while consume extensively more memory.
- Once missing some necessary conditions, *no guaranteed* approaches would fail.
- Our approach learns to *approximately* achieve CG and balances performance and resource usage.

References

[1] Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[2] Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2019.

http://lfzhao.com

Poster Session

zhao.linf@northeastern.edu