Discovering Group Structures via Unitary Representation Learning

Dongsung Huh

MIT-IBM Watson AI Lab

ICLR

Introduction

- Groups are foundational across scientific domains: e.g. Mathematics, Physics, Computer science. Discovering group structures in data is a key challenge.
- Group axioms are non-differentiable, making them hard to integrate into deep learning frameworks.
- We present a differentiable approach to discovering group structures, leveraging representation theory of finite groups.

Groups

A group $(G, \circ)$ is a set G with a binary operation $\circ$ that satisfies four axioms:

Closure:	$\forall a, b \in G,\ a \circ b \in G$
Associativity:	$\forall a, b, c \in G,\ (a \circ b) \circ c = a \circ (b \circ c)$
Identity:	$\exists e \in G,\ \forall g \in G,\ g \circ e = e \circ g = g$
Inverse:	$\forall g \in G,\ \exists g^{-1} \in G,\ g \circ g^{-1} = g^{-1} \circ g = e$

Group Representations

A representation ϱ of a group $(G, \circ)$ on a vector space V is a *group homomorphism* $\varrho\colon G \to \mathrm{GL}(V)$, preserving the group structure:

$$\varrho(g_1 \circ g_2) = \varrho(g_1)\varrho(g_2), \qquad \forall g_1, g_2 \in G.$$

A representation ϱ is called *unitary*, if $\forall g \in G,\ \varrho(g)$ is a unitary transformation, *i.e.* preserving the inner product.

Unitarity Theorem: For compact and finite groups, every finite-dimensional representation is equivalent to a unitary one.

Task: Binary Operation Completion (BOC)

- Complete the Cayley tables of binary operations over discrete symbols S.
- Isolates the fundamental challenge of discovering group structures solely from interactions between elements, eliminating any confounding factors.
- Closely resembles Matrix Completion, which provides a theoretical foundation for numerous applications, e.g. recommender systems, compressed sensing.

Background: Low-rank Matrix Completion

- Classical methods enforce explicit low-rank constraint or minimize the nuclear norm as a convex proxy for rank.
- Matrix factorization, when regularized with L2 regularization (or initialized with small weights), implicitly defines a complexity metric that approximates rank, e.g. nuclear norm or Schatten norm.
- Implicit approaches have demonstrated superior performance in matrix completion, especially with limited data (Arora et al., 2019).

Modeling Framework

Linearize BOC to Tensor Completion Problem:

Symbolic Binary Operation $\Longrightarrow$ Bilinear Map (over a vector space V)

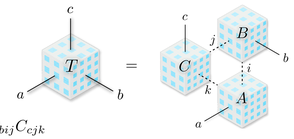
Linearize

$\circ : S \times S \to S \Longrightarrow D : V \times V \to V$ $\qquad D \in \mathbb{C}^{n \times n \times n}$ Sparse Data Tensor

$a \circ b = c \Longrightarrow D_{abc} = \begin{cases} 1 & \text{if } a \circ b = c \\ 0 & \text{otherwise} \end{cases}$ Recover missing entries of D from partial observation

HyperCube Factorization

Train a model tensor T to recover D. Factorize T as a product of three order-3 factors: $A, B, C \in \mathbb{C}^{n \times n \times n}$



$$T_{abc} = \frac{1}{n}\mathrm{Tr}[A_a B_b C_c] = \frac{1}{n}\sum_{ijk} A_{aki} B_{bij} C_{cjk}$$

Combined with a regularizer that promotes unitarity of factor embeddings.

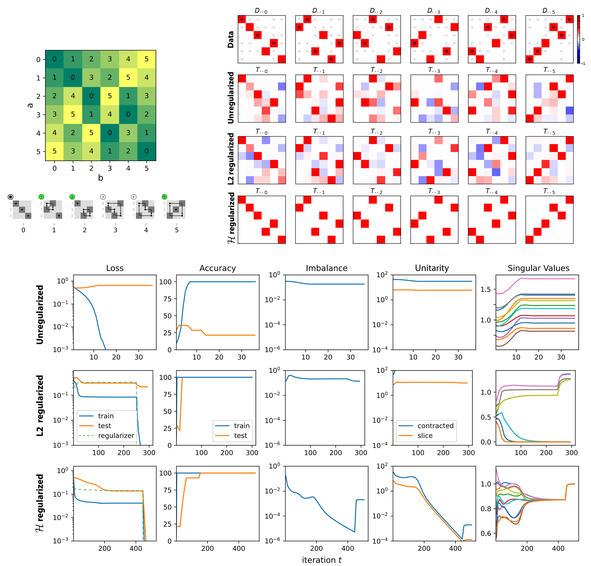
$$\mathcal{H} \equiv \left\|\frac{\partial T}{\partial A}\right\|_F^2 + \left\|\frac{\partial T}{\partial B}\right\|_F^2 + \left\|\frac{\partial T}{\partial C}\right\|_F^2 = \frac{1}{n}\,\mathrm{Tr}\left[A_a^\dagger A_a B_b B_b^\dagger + B_b^\dagger B_b C_c C_c^\dagger + C_c^\dagger C_c A_a A_a^\dagger\right]$$

Key properties grounded on the Representation Theory of Groups:

- Interactions between group elements are modeled as the products of matrix embeddings, inheriting associativity of matrix multiplications
- Leverages **unitarity theorem** to promote solutions within the space of unitary embeddings. This effectively constrains parameters to a relevant subspace without loss of generality: Faster convergence & improved sample complexity.

This factorized architecture and regularizer instills a strong implicit bias towards discovering valid group structures.

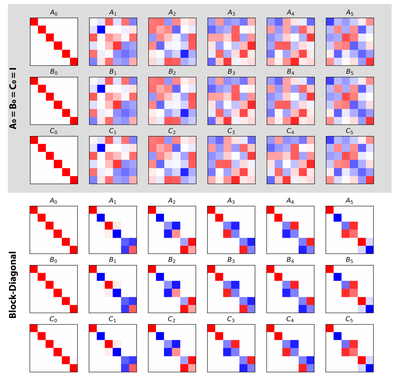
Learning Dynamics on S3 Task



- HyperCube regularization causes all singular values of embeddings to collapse to one: i.e. unitary (full-rank) matrices. Exact recovery of the data tensor D.
- In contrast, L2 regularization yields low-rank embedding. Imperfect recovery of D.
- This bias towards unitarity is highly unusual among deep learning models.

Learned Embeddings are Unitary Representations

- Learned matrix embeddings are unitary
- Exhibits shared embedding $A_g = B_g = C_g^\dagger$
- And group homomorphism $A_{g_1} A_{g_2} = A_{g_1 \circ g_2}$
- Thus, they form proper unitary matrix representation: $\varrho(g) = A_g = B_g = C_g^\dagger$
- In fact, the regular representation of group: $\mathrm{Tr}[\varrho(g)] = \begin{cases} n & \text{if } g = e, \\ 0 & \text{otherwise.} \end{cases}$
- Contains the complete set of irreducible representations.



Key Operating Mechanism: Combing the above results yields

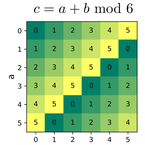
$$T_{abc} = \frac{1}{n}\mathrm{Tr}[\varrho(a)\varrho(b)\varrho(c)^\dagger] = \frac{1}{n}\mathrm{Tr}[\varrho(a \circ b \circ c^{-1})] = D_{abc}$$

which indeed produces the data tensor *D*. This mechanism **universally applies to all groups**.

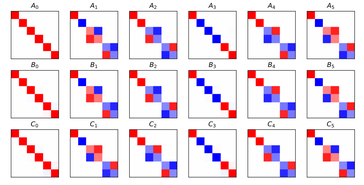
Small-scale BOC Experiments

Group:

$c = a + b \bmod 6$

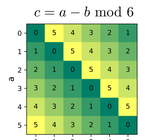


$A_g = B_g = C_g^\dagger = \varrho(g)$

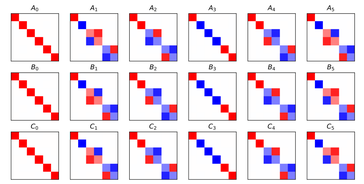


Non-associative:

$c = a - b \bmod 6$

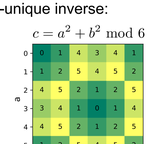


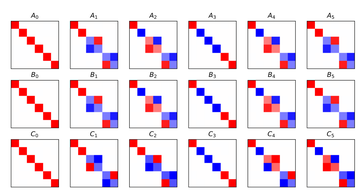
$A_g^\dagger = B_g = C_g = \varrho(g)$



Non-unique inverse:

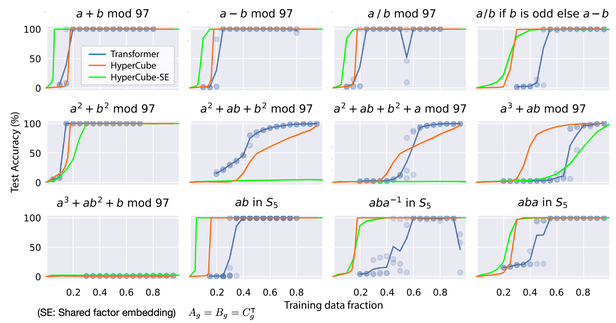
$c = a^2 + b^2 \bmod 6$



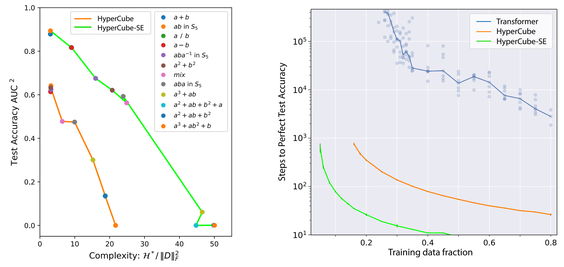


Large-scale BOC Experiments

- Large task dataset from [Power et al 2022].
- HyperCube shows clear priority towards learning simple (group and group-like) operations, requiring more data to learn complex operations. Shared-embedding (HyperCube-SE) $A_g = B_g = C_g^\dagger$ further sharpens the bias towards groups.
- Transformers also exhibit a form of simplicity bias, but not towards groups.



Complexity of an operation can be implicitly defined as the minimum regularizer loss $\mathcal{H}^*$ when fitting the fully observed operation D (i.e., under constraint $T = D$).



Left: Generalization trend shown as a function of complexity, revealing a clear monotonic relationship: as task complexity increases, generalizability decreases (i.e. total area under the test accuracy curve).

Right: HyperCube exhibits exceptional learning speed, **100X** faster than Transformer baseline in the S5 task, eliminating the "grokking" phenomenon. HyperCube-SE achieves an additional **10X** speedup and requires only 5% of data for perfect accuracy.

Discussion

HyperCube exhibits universal inductive bias towards general algebraic structure of groups, with implication for **automatic symmetry discovery**.

Scalability: HyperCube's compute/memory cost scales as O(n^3), but it can be reduced to O(n^2) by utilizing band-diagonal embeddings.

Open Questions: Deriving rigorous generalization bounds for BOC tasks and formally proving the optimality of unitary representations remain as open problems. Furthermore, extending our methodology to accommodate multiple symbols and operations beyond binary operations would significantly broaden its scope and applicability.