

UCLA

UCLA Electronic Theses and Dissertations

Title

Inspecting Generalization of Reinforced Learners: The HALMA Benchmark

Permalink

<https://escholarship.org/uc/item/9935j5g4>

Author

Ma, Xiaojian

Publication Date

2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Inspecting Generalization of Reinforced Learners:
The HALMA Benchmark

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Xiaojian Ma

2020

© Copyright by

Xiaojian Ma

2020

ABSTRACT OF THE THESIS

Inspecting Generalization of Reinforced Learners:
The HALMA Benchmark

by

Xiaojian Ma

Master of Science in Computer Science

University of California, Los Angeles, 2020

Professor Song-Chun Zhu, Chair

Humans learn compositional and causal abstraction, *i.e.*, knowledge, in response to the structure of naturalistic tasks. When presented with a problem-solving task involving some objects, toddlers would first interact with these objects to reckon what they are and what can be done with them. Leveraging these concepts, they could understand the internal structure of this task, without seeing all of the problem instances. Remarkably, they further build cognitively executable strategies to *rapidly* solve novel problems. To empower a learning agent with similar capability, we argue there shall be three levels of generalization in how an agent represents its knowledge: perceptual, conceptual, and algorithmic. In this work, we devise the very first systematic benchmark that offers joint evaluation covering all three levels. This benchmark is centered around a novel task domain, HALMA, for visual concept development and rapid problem solving. We conduct extensive experiments on reinforcement learning agents with various inductive biases and carefully report their proficiency and weakness.

The thesis of Xiaojian Ma is approved.

Lin F. Yang

Kai-Wei Chang

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2020

To my parents and friends.

TABLE OF CONTENTS

1	Introduction	1
1.1	Related Work	3
2	Background	5
2.1	Three-level Generalization	5
2.2	Markov Decision Processes	7
3	HALMA: Humanlike Abstraction Learning Meet Affordance	9
3.1	The HALMA Domain	9
3.2	The Concept Space of HALMA	11
3.3	Task Formulation and Evaluation	12
3.4	Generalization Test	13
4	Benchmarking RL Agents with HALMA	16
4.1	A Collection of RL Learner	16
4.2	Experiment 1: Generalization Tests	18
4.3	Experiment 2: Ablation Studies	21
4.3.1	Ablation on the Volume of Training Set	21
4.3.2	Ablation on the Maximum Option Length	23
5	Conclusive Remarks	26
	References	27

LIST OF FIGURES

3.1	(a) Given a visual panel with various colored MNIST digits and a hint, an autonomous agent is tasked to reach the goal in a maze. The concept space guides the generation of the visual panels; it consists of (b) spatial grammar, (c) temporal grammar, and (d) causal structure. (e) The semantics and affordance of the colored MNIST digits are augmented on the corresponding maze; the maze is not shown to the agent.	9
4.1	Architecture of the actor model, where T is equal to <code>max_opt_len</code>	18
4.2	Architecture of the critic model, where T is equal to <code>max_opt_len</code>	18
4.3	Ablation study of different number of training mazes.	22
4.4	Ablation study of different <code>max_opt_len</code> (symbolic observations).	24
4.5	Ablation study of different <code>max_opt_len</code> (visual observations).	25

LIST OF TABLES

4.1	Architectural parameters of evaluated agents	17
4.2	Hyper-parameters of TD3	19
4.3	Examples and results of generalization tests (- indicates no problem is dynamically generated)	21

ACKNOWLEDGMENTS

I would like to send my sincere gratitude towards Professor Song-Chun Zhu, my advisor for his continuous guidance and support on my research during the graduate study. Specifically, I am grateful for the opportunity he offered to work with the exceptionally talented colleagues at the VCLA lab. I would also like to thanks all the lab members, in particular, Sirui Xie and Yixin Zhu for their generous help and mentorship on how to become a good researcher.

The thesis consists of a conference submission, titled as "HALMA: Humanlike Abstraction Learning Meets Affordance in Rapid Problem Solving" which was submitted to 2021 International Conference on Learning Representation (ICLR) [Ano21]. The list of coauthors is Sirui Xie, Xiaojian Ma, Peiyu Yu, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. The work reported herein was supported by ONR MURI grant N00014-16-1-2007, ONR N00014-19-1-2153, and DARPA XAI N66001-17-2-4029.

CHAPTER 1

Introduction

Learning to generalize across varied environments and task is arguably the central quest for modern intelligent agents. Recently, researchers with different disciplines including AI, cognitive science, psychology and neuroscience have been approaching to a general consent that the emergence of generalizable concepts organized by proper language-like processing mechanisms seems to be the key [GFP19, MBN10, MHR20, Gri20, GB20]. While the evolution of human intelligence also suggests that these architectures should not be merely hand-crafted. Rather, some bootstrapping with scarce supervision and innately specified predispositions would primarily capture the dynamics [Kar94, AGP15, Cho86]. Therefore, the essential question is: What is the proper machinery to learn these generalizable concepts from scarce supervisions? By *scarce supervision*, we mean the way to provide supervision is akin to how you teach Ada; one only provides sparse and indirect feedback without direct rules or dense annotations. By *generalizable concepts*, we emphasize more than the competence of memorization and interpolation; the learned representation ought to appropriately extrapolate and generalize in out-of-distribution scenarios. Such a superb generalization capability is often regarded as one of the celebrated signatures of human intelligence [LST15, Mar18, LB18]; it is attributed to rich *compositional* and *casual* structures in human mind [FP88].

Inspired by these observations, in this work, we quest for a computational framework to learn abstract concepts emerged in challenging and *interactive* problem-solving tasks, with a humanlike generalization capability: The learned abstract knowledge should be easily transferred to out-of-distribution scenarios.

The general context of interactive problem solving poses extra challenges over classic settings of concept learning; instead of merely emerging concepts, it further demands the learning agent to leverage such emerged concepts for decision-making and planning. Ada, after understanding semantics and affordance in *Halma*, can effortlessly perceive and parse novel scenarios [ZGF20]. Yet, she would still struggle in strategically playing the game as she needs to decide among multiple affordable moves. In essence, the central question is: If conceptual knowledge can generalize as such, what meta-benefits does it offer on solving unseen problems [SZW96]?

The classic decision-making account of these meta-benefits would be: Leveraging knowledge, we can develop cognitively executable *strategies* with high planning [San08] and exploration efficiency [KLC98]; these strategies facilitate us to solve problems rapidly in unseen scenarios. They are what we call the *algorithms* or *heuristics* of this task.

Taking a step further, [WKK18, GMG19] hypothesize that modern reinforcement learning agents, incentivized by these meta-benefits, have already discovered such algorithms. However, to date, their argument is still speculative since these agents have not been evaluated in tasks with rich internal structures yet limited exposure [LUT17, KSM17]. A diagnosis benchmark for generalization capability is thus in demand to bridge communities of concept development and decision-making.

The main contribution of this work is a *Halma*-inspired competence benchmark: Humanlike Abstraction Learning Meets Affordance (HALMA). We rigorously devise HALMA with three levels of generalization in visual concept development and rapid problem solving; see details in Section 2.1. HALMA is unique in its *minimum yet complete* concept spaces, a miniature of compositional and causal structures in human knowledge. It *dynamically* generates test problems to informatively evaluate learning agents' capability in out-of-distribution scenarios *under limited exposure*. We conduct extensive experiments with reinforcement learning agents to benchmark proficiency and weakness.

1.1 Related Work

Recently, there emerges a burst-out of benchmarks for diagnosing a set of clearly defined competencies of AI systems, which we draw inspiration from and sincerely honor. In a word, HALMA differentiates from all of them in its holistic evaluation towards all three levels of generalization.

Readers may be curious about the relation between HALMA and conventional navigation tasks such as [MPV17]. We hope we have made it clear the difference between HALMA and them in 3.1 of main text: In these navigation tasks, there is only one maze, and new problem instances are simply new combinations of initial and goal states. Hence, rapid problem solving only requires agents to memorize the whole maze, whereas in HALMA the only shared structure between problem instances is the concept space. Going beyond memorization, HALMA requires two extra cognitive abilities—understanding and reasoning. We also notice that in another embodied navigation task, the Habitat challenge [SKM19], agents are indeed evaluated in completely unseen environments, under the protocol of which [WKM20] has achieved close-to-optimal performance with large-scale training. However, without a clearly specified concept space, the evaluation in Habitat is akin to the Random Split in HALMA under the setup of `max_opt_len=1`. The reason why we emphasize `max_opt_len` is that the very idea of *affordance* is only interesting if the action/option space is large enough and highly structured. Otherwise, when `max_opt_len=1`, agents with memory or attention do generalize well in both Random Split and our Dynamic Test; see detailed results in 4.3.2. Perhaps the notion of *affordance* seems a bit abstract in HALMA and can be more intuitive in visual semantic navigation and control [YWF19, CLS20]. We hope our work can inspire the future development of benchmarks for these topics.

Compositional Language and Elementary Visual Reasoning (CLEVR) [JHM17] is one of the earliest datasets that diagnose models’ visual reasoning abilities. High-level reasoning skills required in CLEVR include counting, comparing, logical inference, and memory. The

same set of skills are also required in HALMA, but without the guidance of language. Accounting for a similar purpose, [BMN19] propose a minimalist alternative, Spatial Queries On Object Pairs (SQOOP). While relations in SQOOP are only spatial, benchmarks inspired by Raven’s Progressive Matrices (RPM) are proposed towards abstract visual reasoning [BHS18, ZGJ19], in which the capacity of sequential decision making is not required. In sum, all prior works listed in this paragraph are discriminative tasks. Different from them, the generative nature of interactive problem solving in HALMA is akin to human exploration in the open-ended world.

As for planning and reinforcement learning, Box-World and StarCraft II minigames [VEB17] in [ZRS19] are tasks that also require relational concept learning; the concepts within, however, are mostly spatial.

In contrast, the concept space in HALMA is abstract and complex. The mapping from the visual space to the semantic space is non-trivial to learn, which requires agents’ understanding of the temporal grammar and the causal structure. Moreover, HALMA is a partially observable domain that requires dedicated efforts for exploration.

The closest one that is also inherently generative, compositional, and abstract is the Simplified version of the CommAI Navigation (SCAN) [LB18], an instruction following task. Essentially, SCAN is seq2seq translation, with little uncertainty or variation in primitives. Hence, it does not test agents’ perceptual generalization or algorithmic generalization. In contrast, HALMA is a task for visual concept development and rapid problem solving. Agents need to understand concepts from visuomotor experience and make smart decisions to acquire utility.

CHAPTER 2

Background

2.1 Three-level Generalization

Our motivations might seem, *prima facie*, bold. To convince readers and support our optimism, we summarize some recent progress in this section. In particular, we provide a taxonomy of three levels of generalization on a competency basis. Indeed, generalization is a multifaceted phenomenon. Previous evaluations for generalization were predominantly defined in a statistical sense, following the classical paradigm of train-evaluation-test random split [CKH19] while ignoring internal structures. However, we argue this classical paradigm should not be the only objective approach wherein agents can or should generalize beyond their experience [BHS18], especially if our goal is to construct humanlike general-purpose problem-solving agents [LUT17].

Perceptual Generalization Perceptual generalization characterizes agents’ capability to represent unseen perceptual signals, *e.g.*, *appearance* or *geometry* in vision. In his seminal book, *Vision*, [Mar82] describes the process of vision as constructing a set of representations, parsing visual sensory data into descriptions. Such descriptions provide *conceptual primitives* [Car09] for agents’ understanding of the environment, boosting the efficacy of downstream cognitive activities (*e.g.*, memory, learning, and reasoning). Learning an object-oriented representation of independent generative factors without supervision is thus believed to be a crucial precursor for the development of humanlike artificial intelligence. Although unsupervised disentanglement and segmentation [EHW16, HMP17] resurged years ago, it is only till [LBL19] did we realize the importance of evaluation on their generalization. More recently,

[BMW19], [GKK19], and [LWP20] evaluate their disentanglement/segmentation models outside of training regimes, especially on unseen combinations of visual attributes and numbers of objects.

Although a hypothetically perfect *semantic* description can truthfully represent the primitive concept of “what it is,” it could only contribute partially to achieving the understanding of “what can be done with it” [MLB08, ZZZ15]. Humanlike agents should equip with such task-oriented abstraction, *affordance*, supported by compelling evidences in the field of developmental psychology; for instance, 18 to 24-month-old infants can distinguish *bootstrapped concepts* [Qui60], such as “a walkable step is not a cliff” [KA13].

At a computational level, given a task specified by a Markov decision process, irrelevant features should be *abstracted out* [LWL06, FPP11, KAC20]. Representation learned in this way bootstraps conceptual content. Recently, disentanglement as such has demonstrated efficacy [GKB19, WHA18] and elementary perceptual generalizability [ZMC20].

Conceptual Generalization While perceptual generalization closely interweaves with vision and control, conceptual generalization resides completely in cognition, assuming the readiness of all primitive concepts and some bootstrapped ones. The central challenge in conceptual generalization¹ is: How well can an agent perform in unseen scenarios given *limited exposure* to the underlying *configurations* [Gre93]? It is connected with the Language of Thought Hypothesis [FP88, GTF08]: The productivity, systematicity, and inferential coherence in languages characterize compositional and causal generalization of concepts [LST15].

How to learn representations with conceptual generalization is still an open question, drawing increasing attention in our community. With a synthetic translation task, [LB18] reveal the incompetence of general purpose recurrent models [Elm90, HS97, CGC14] in generalizing to (i) unseen primitives, (ii) unseen compositions, and (iii) longer sequences than training data. Similar incompetence of relational inductive biases [BHB18] on hard composi-

¹Conventionally, it is dubbed *combinatorial* generalization or *systematic* generalization. We use the term *conceptual* to highlight its functional signature.

tional extrapolation has also been exemplified in abstract visual reasoning [BHS18]. Notably, there is also a line of research on *emerging* these linguistic structures from bootstrapped communication [LHT18, MA18].

Algorithmic Generalization Agents’ understanding of the structured environment should be reflected in their performance in solving novel problem instances; they ought to build strategies upon the developed concepts, resembling *cognitive control* in human mind [RNB05, BC14]. We use the term algorithmic generalization to describe such flexibility. Specifically, for a problem domain where the internal structure contains an optimal exploration strategy, algorithmic generalization requires agents to discover this optimal strategy to explore efficiently in *new* problem instances. For example, in the domain of dependent bandit problems designed by [WKT16], there is one arm whose return leaks the index of the optimal arm. Given a new problem, agents who discovered the algorithm of this domain would first try the leaky arm and then go straight to the optimal arm. Furthermore, as an acid test, algorithmic generalization also measures the agent’s ability in long-term planning in unseen problem configurations, after acquiring adequate information. Evaluation as such has been discussed by [TWT16] and [GMG19].

Problem domains discussed above, however, still lack rich concept spaces, nor do they test agents’ perceptual generalization, omitting the interaction among the three levels introduced in this paper. Essentially, they are still far-off from the famous Atari game, Frostbite, which is argued to be a testbed for humanlike problem solving [LUT17]. In this work, we introduce a new problem domain to facilitate joint efforts towards representations with these three levels of generalization.

2.2 Markov Decision Processes

For modeling the action decision process in our context, a standard Markov decision process (MDP) [SB18] $(\mathcal{S}, \mathcal{A}, r, \mathcal{T}, \mu, \gamma)$ is considered, where \mathcal{S} and \mathcal{A} denotes the space of feasible states and actions respectively, $r(s, a) \rightarrow \mathbb{R}$ is the reward function, $\mathcal{T}(s'|s, a)$ and $\mu(s)$

represent the transition probability and initial state distribution and $\gamma \in (0, 1)$ is the discount factor. A stochastic policy $\pi(a|s) : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ maps state into action distribution. A trajectory ζ is given by the sequence of state-action pairs $\{(s_0, a_0), (s_1, a_1), \dots\}$.

CHAPTER 3

HALMA: Humanlike Abstraction Learning Meet Affordance

3.1 The HALMA Domain

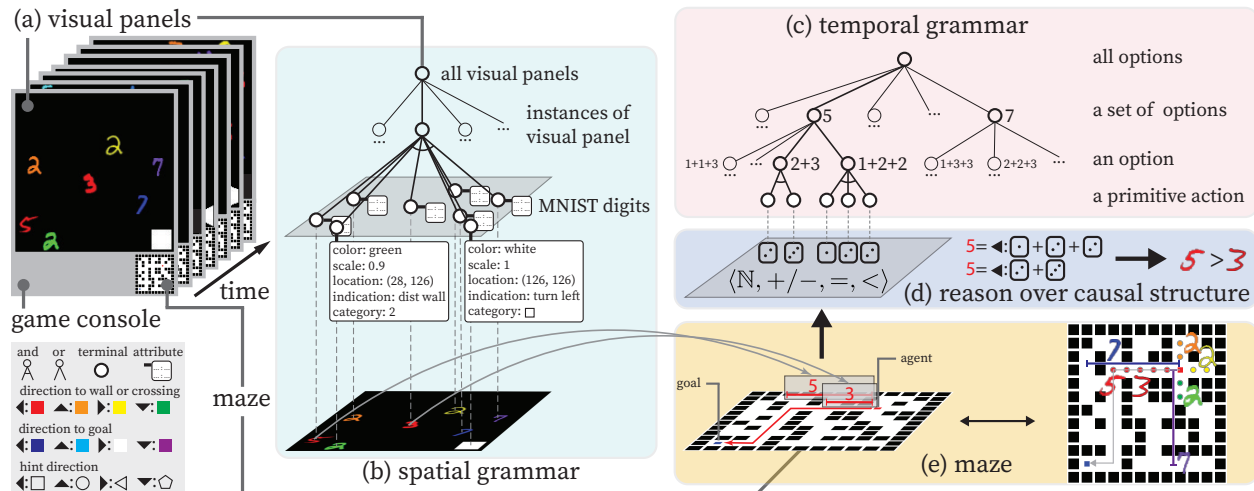


Figure 3.1: (a) Given a visual panel with various colored MNIST digits and a hint, an autonomous agent is tasked to reach the goal in a maze. The concept space guides the generation of the visual panels; it consists of (b) spatial grammar, (c) temporal grammar, and (d) causal structure. (e) The semantics and affordance of the colored MNIST digits are augmented on the corresponding maze; the maze is not shown to the agent.

The setup of HALMA is minimal and interpretable. Instead of replicating the entire game of *Halma*, we only preserve the most essential ingredients: The learning agent is cast as one pawn, navigating around the “magical” *Halma* landscape by itself. To simplify the environment without loss of generality, we build a maze in a grid-world for each *scenario* (or *problem* henceforth), resembling a *cognitive map* of the agent. Distinct from vanilla grid-world maze games, HALMA is novel in terms of our design of its observation space and

action space. The agent perceives neither the global map nor any local patch of the global map; instead, it is shown with a visual panel of various numbers of MNIST digits in various color, randomly scaled and placed; see 3.1 (a). These colored digits indicate the *semantics* of (i) the distance till a wall towards each direction, (ii) the distance till the nearest crossing or T-junction towards each direction, and (iii) the distance and direction to the goal; the visual panel only displays non-zero distances. For example, in 3.1 (a) (e), **5** indicates the wall to the left is 5-grid away, and **3** indicates the nearest crossing is 3-grid away to the left; the visual color of **red** refers to the semantics of “left.” The agent will also be hinted with a symbol from the set $\{\circ, \triangle, \square, \diamond\}$ at any crossing for the correct direction; see an example of \square in 3.1 (a). When making a decision, the agent needs to first select a direction and then select either a primitive action or an option composed by a sequence of primitive actions [SPS99] with maximum length `max_opt_len`. The direction set is $\{\blacktriangle, \blacktriangledown, \blacktriangleleft, \blacktriangleright\}$. The primitive action set, in terms of the number of moves, is $\{\square, \square, \square, \square\}$; this design of primitive numbers with a maximum of three aligns with the doctrine of core knowledge in developmental psychology [FC03, Deh11]. If an option is selected, consecutive hops as in *Halma* are simulated; all observations from intermediate states will be skipped, and only the observation of the final state is provided. A move would fail if a wall stops the agent, leaving the agent’s position unchanged; failure moves bring penalties to the agent. The agent would receive a positive reward when reaching the goal. Such a design encourages the agent to comprehend which MNIST digit *affords* it to take which moves.


Essentially, HALMA is a 2D contextual navigation game, sharing the same spirit with those in [MPV17] and [RWK18]. However, *contexts* in these prior works are elusive and conceptually meaningless. As such, they only evaluate generalization at either the visuomotor or algorithmic level. In stark contrast, HALMA is unique, possessing a rich, crisp, and challenging configuration space of problems, semantics, and affordance; see details in the next subsection.

3.2 The Concept Space of HALMA

Producing visual panels heavily relies on the concept space. The concept space of HALMA consists of an explicit *spatial grammar* for visual panels, an implicit *temporal grammar* for actions and options, and an underlying *causal structure* that specifies the intersection of spatial and temporal grammar. For simplicity, we only introduce them verbally here; see an illustration in 3.1. Intuitively, the spatial grammar produces all possible descriptions of visual panels, spanning all configurations of *semantics* introduced in 3.1. To generate a visual panel for a given state, we first sample an MNIST digit for each entry of its description and then sample a random scale and position. The sampled MNIST digit is then colored on the basis of its semantics, *i.e.*, directions to a wall, a crossing, or a goal; see 3.1 (b) and the legend. The temporal grammar produces all possible moves, either a single primitive action or a composed option, regardless of the visual stimuli. For instance, a non-terminal node $\blacktriangleleft : 5$ can be parsed into options **opt**, such as $\blacktriangleleft : \square + \square + \square$ and $\blacktriangleleft : \square + \square$; see 3.1 (c). Despite of their distinction in terms of how an option is decomposed into primitive actions, these options are equivalent in their causal effects. Specifically, these causal effects bind visual MNIST digits with digital actions based on one of the simplest mathematical structures in human cognition [Fla63]: $\langle \mathbb{N}, +/-, =, < \rangle$; namely, *natural numbers* \mathbb{N} , *operations* $+/-$, and *relations* $=, <$ over \mathbb{N} . For example (see also 3.1 (d)), a learning agent is expected to understand relations between **5** and **3** via

- $\langle \mathbf{S}, < \rangle$: the set of semantic *generators*¹ with an *order* over it, *e.g.*, **3** < **5**;
- $\langle \mathbf{A}, +/-, = \rangle$: the set of affordance generators with operations and equality, *e.g.*, **5** = $\blacktriangleleft : \square + \square + \square = \dots$;
 $\square + \square + \square = \blacktriangleleft : \square + \square = \dots$;
- $\langle \mathbf{A}, +/-, < \rangle$: the set of affordance generators with operations and inequality, *e.g.*, $\blacktriangleleft : \square + \square < \mathbf{5}$, **5** < $\blacktriangleleft : \square + \square + \square$;

¹For the sake of formalism, we adopt the terminology from General Pattern Theory [Gre93], wherein the term *generator* refers to basic units in a *configuration space*. Intuitively, an object file [KTG92], is a semantic generator. It is also a generator for configuration spaces of affordance and causality, for which actions/options are also generators.

- $\langle \mathbf{C}, +/-, = \rangle$: the set of causal generators with operations and equality, *e.g.*, $\mathbf{5} = \mathbf{3} + \mathbf{2}$: .

3.3 Task Formulation and Evaluation

We expect agents who developed the concept space to leverage this knowledge and rapidly solve new problems in HALMA. To this end, we formulate this rapid problem-solving task with an objective to *maximize the agent’s rewards accumulated over a few trials in a novel problem instance*:

$$\mathbb{E}_{\zeta} \left[\sum_{i=0}^N \gamma^{\sum_{j=0}^{i-1} \text{len}(\tau_j)} \sum_{t=0}^{\text{len}(\tau_i)-1} \gamma^t R(s_{\tau_i,t}, a_{\tau_i,t}) \right]. \quad (3.1)$$

Specifically, an agent’s experience in each problem instance is dubbed an *episode* ζ [WKT16], which terminates when a maximum number of *steps* L is reached or a maximum number of *trials* N have been accomplished. A *trial* τ proceeds with actions $a_{\tau,t}$, spanning multiple *steps* t ; it starts from an initial state s_0 and terminates when the agent reaches the goal s_g (thus accomplished), or when it consumes the maximum number of steps H (thus failed). The agent is respawned to the initial state when a trial terminates. It is awarded $R(s_g, \cdot)$ if the trial is accomplished. The cumulative reward in one episode is the sum of temporally γ decayed accomplishments. When one episode terminates, the agent is presented with the next problem.

Under this task formulation, learning agents should be evaluated against oracle solutions, analogous to ground-truth annotations in supervised learning; recall that the oracle agent has complete understanding of the concept space and the problem domain. Since HALMA is a *partially observable* domain, its oracle behavior consists of two aspects: optimal exploration and optimal planning. As introduced in 3.2, problems are generated by adding deceptive branches to optimal paths. Hence, the optimal exploration strategy is to stop at each crossing to obtain the hint from the visual panel. Intuitively, the agent should understand “when two digits with the same color are exhibited in the visual panel, the *lesser* one indicates the crossing, and I should stop there for hint” based on the concept of $\langle \mathbf{S}, < \rangle \cup \langle \mathbf{A}, +/-, < \rangle$. An oracle agent would sacrifice the first trial to explore; note that the cost is still low as it would

explore along the optimal path with the guidance of hints, avoiding all deceptive branches. Afterwards, the oracle agent should retrieve its experience and merges consecutive moves towards the same direction to form the optimal plan. Take the maze example shown in 3.1 (e); during exploration, the agent sees a **5** and a **3** in the visual panel and takes an option $\blacktriangleleft : \square + \square$ to obtain a hint \square , which guides it to keep moving left $\blacktriangleleft : \square$ until the wall. Then in the second trial, the agent should exploit $\langle \mathbf{A}, +/-, = \rangle \cup \langle \mathbf{C}, +/-, = \rangle$ via $\blacktriangleleft : \square + \square + \square$. With this oracle agent, we can have evaluation metrics normalized across different problems. Instead of directly calculating the ratio of 3.1 between proposed agents and the oracle agent, which involves strong non-linearity, we carefully decompose it into three metrics with more intuitive measures:

- Ratio of invalid moves $\rho_a = \mathbb{E}_\zeta \left[\frac{\# \text{invalid moves}}{\sum_i \text{len}(\tau_i)} \right]$ for semantics and affordance understanding;
- Success rate of goal reaching $\rho_g = \mathbb{E}_\zeta \left[\frac{1}{N} \sum_i \delta(s_{\tau_i, -1} = s_g) \right]$ for leveraging concepts to explore;
- Efficiency in exploration and planning $\rho_p = \mathbb{E}_\zeta \left[\frac{1}{N} \sum_i \frac{\text{len}(\tau^*)}{\text{len}(\tau_i)} \right]$ for algorithmic understanding.

3.4 Generalization Test

One of our key contributions in HALMA is a novel paradigm to test agents’ capability in all three levels of generalization, which extends the classical paradigm of statistical learning. Our training set consists of 100 mazes² along with their visual panels. Different from the classic paradigm, the evaluation of agent’s performance in HALMA would emphasize on the *explicit extrapolation* test, which should be conducted in the *held-out* compositional and relational configurations; such design echoes recent trend in evaluating agent’s generalization capability [BMW19, LB18, ZRS19]. Compared to these prior domains, HALMA is unique as it is a partially observable and interactive problem-solving task, wherein an agent is tasked to *autonomously* learn the immense concept space and form the abstract knowledge. Hence,

²This design reflects our thesis argument, *i.e.*, agents shall generalize their understanding from limited exposure to the concept space. An ablation study on the volume of training set can be found in 4.3.1.

simply holding off a *pre-selected, fixed* subset of conceptual configurations would impose severe restrictions on problem generators. For instance, if we would like to allow agents to see a **4**, they must be able to see a **3** by simply moving $\blacktriangleleft : \square$ from where they see **4**. In other words, if we managed to strictly withhold **3** from agents, they would not see any red digits larger than 3 in this *interactive* problem solving task. Therefore, an *ex post* evaluation protocol that *dynamically* generates tests is more desirable.

In this paper, we propose an ingenious solution: Instead of *aimlessly* generating a large test set of *random* cases, we devise an algorithm to *proactively* generate *tailored* tests in accord to what the agent might have learned; this design would produce a definitive and much more informative evaluation of agent’s competence. The intuition is simple: When a teacher finds a student consistently make right decisions during training, wherein the student only needs to understand **3** < **5** and **4** = **2** + $\blacktriangleleft : \square$, the teacher may quiz the student on **3** vs **5** and **2** vs **4**. To implement this protocol in HALMA, we first store agents’ experience during training as their external memory MEM. We then construct a representation to emulate agents’ *knowledge bases* (KB) for $\langle \mathbf{S}, < \rangle$ and $\langle \mathbf{A}, +/ -, = \rangle \cup \langle \mathbf{C}, +/ -, = \rangle$: $\text{KB}_{\mathbf{S}}$ tracks the agent’s understood configurations on semantics, and $\text{KB}_{\mathbf{A} \vee \mathbf{C}}$ tracks the agent’s understood configurations on affordance and causality. Here, we assume that (i) valid decisions³ in experience were made upon understanding *inequality* configurations, and (ii) agents understand configurations involving *equality* and *operations* in experienced transitions. With these KBs, we *dynamically* generate test problems with novel configurations, wherein agents should likewise act appropriately if they understood not only seen configurations but their underlying concepts.

Tests in HALMA are on the competence basis: Conceptual generalization is built upon perceptual generalization, with the algorithmic generalization resides on top. Tests for perceptual generalization are backed by the *spatial grammar*, including unseen MNIST images

³Note that some decisions may come from random exploration. We introduce a threshold on the visitation count to filter them out.

and unseen compositions of visual attributes, *i.e.*, shape and color. Tests for conceptual generalization are based on *the concept of* $\langle \mathbb{N}, +/-, =, < \rangle$, consisting of novel *equality* and *inequality* configurations. Results of these two tests are manifested in algorithmic generalization. Specifically, agents could only pass all of these tests by making right *exploration* decisions based on relations of novel digit pairs $\langle \mathbf{d}_1, \mathbf{d}_2 | \mathbf{type} \rangle$, where **type** refers to various directions. Inappropriate exploration may cause agent to miss hints at crossings or to be trapped in dead-ends, resulting in failures of the tests. Moreover, these novel digit pairs also test the agents' understanding of the *temporal grammar*, requiring agents to make proper *exploitation* decisions by merging novel consecutive actions/options into a *greater* option.

Since conceptual generalization connects the other two, all three levels of generalization are covered when test problems are dynamically generated with novel configurations in $\langle \mathbb{N}, +/-, =, < \rangle$. Recall that the generation mechanism of a problem is to first generate an unseen configuration of optimal path and then add deceptive branches; the latter is pivotal for a test problem since it involves generating novel digit pairs $\langle \mathbf{d}_1, \mathbf{d}_2 | \mathbf{type} \rangle$. By design, the lesser digit within a pair should indicate the distance to the nearest crossing, and the greater the distance to the wall. Hence, agents could be tested by these novel digit pairs, queried based on the agent's KBs. We categorize the problems into:

- Semantic Test (ST): $\text{KB}_{\text{ST}} = (\langle \mathbf{d}_1, \mathbf{d}_2 | \mathbf{type} \rangle \notin \text{KB}_{\text{S}}) \wedge (\exists_{\mathbf{x}} \langle \mathbf{d}_1, \mathbf{d}_2 | \mathbf{x} \rangle \in \text{KB}_{\text{S}})$, *i.e.*, testing visual panels differentiated from KB_{S} in terms of color, shape, or other MNIST digits.
- Affordance Test (AfT): $\text{KB}_{\text{AfT}} = (\forall_{\mathbf{x}} \langle \mathbf{d}_1, \mathbf{d}_2 | \mathbf{x} \rangle \notin \text{KB}_{\text{S}}) \wedge ((\exists \langle \mathbf{d}_1, \mathbf{d}_2 | \mathbf{x} \rangle \in \text{KB}_{\text{A}\vee\text{C}}) \vee (\mathbf{d}_1 = \text{opt}_1 \in \text{KB}_{\text{A}\vee\text{C}} \wedge \mathbf{d}_2 = \text{opt}_2 \in \text{KB}_{\text{A}\vee\text{C}}))$, *i.e.*, testing inequalities inferred from equalities in $\text{KB}_{\text{A}\vee\text{C}}$. **opt** denotes actions or options.
- Analogy Test (AnT): $\text{KB}_{\text{AnT}} = (\forall_{\mathbf{x}} \langle \mathbf{d}_1, \mathbf{d}_3 | \mathbf{x} \rangle \notin \text{KB}_{\text{ST}\vee\text{AfT}}) \wedge (\exists \{ \langle \mathbf{d}_1, \mathbf{d}_2 | \mathbf{x} \rangle, \langle \mathbf{d}_2, \mathbf{d}_3 | \mathbf{x} \rangle \} \subset \text{KB}_{\text{ST}\vee\text{AfT}}) \wedge (\exists \{ \langle \mathbf{d}'_1, \mathbf{d}'_2 | \mathbf{x} \rangle, \langle \mathbf{d}'_2, \mathbf{d}'_3 | \mathbf{x} \rangle, \langle \mathbf{d}'_1, \mathbf{d}'_3 | \mathbf{x} \rangle \} \subset \text{KB}_{\text{ST}\vee\text{AfT}})$, *i.e.*, testing inequalities inferred from the *transitivity* of $<$. $\text{KB}_{\text{ST}\vee\text{AfT}} = \text{KB}_{\text{ST}} \cup \text{KB}_{\text{AfT}}$.

CHAPTER 4

Benchmarking RL Agents with HALMA

4.1 A Collection of RL Learner

The motivating questions of our experiments are: (i) Do model-free agents, exploiting generic inductive biases, develop concepts that generalize in a way, akin to human knowledge? (ii) If there are indeed certain meta-benefits induced by these architectural priors towards problem solving, are they achievable with only limited exposure to the concept space? As it is logistically challenging to experiment with all existing models, a representative subset is culled for benchmark: model-free reinforcement learning agents [WKT16, ZRS19] with gated memory mechanism [HS97], self-attention mechanism [VSP17], or both. Notably, [WKT16] argued that when an RNN agent is fed with previous actions and rewards, its LSTM module would emulate an inner reinforcement learning algorithm; the agent is thus learning to reinforcement learn. They demonstrated that the learned exploration strategy is more efficient than a near-optimal model-free exploration algorithm. [ZRS19] argued that by exploiting stacked attention modules, Transformer agents can conduct iterated reasoning with seen relational units and generalize to unseen scenarios.

We provide a summary, along with some other hyperparameters of all the considered RL learners, in 4.1. The overall model architecture of the agent is then delineated in 4.1 and 4.2.

Table 4.1: Architectural parameters of evaluated agents

Agent	Architecture
<i>Shared</i> Nonlinearity	ReLU
<i>MLP Agent</i> Encoder Decoder	MLP with hidden units [128, 128]. None
<i>LSTM Agent</i> Encoder Decoder	MLP with hidden units [128, 128]. LSTM with layer normalization [BKH16] and hidden units [128].
<i>Transformer Agent</i> Encoder Decoder	A stack of four multi-head self-attention layers, with hidden units [128], four heads and layer normalization, followed by a maximum pooling layer. Parameters are shared across all the attention layers [ZRS19]. MLP with hidden units [128, 128].
<i>Transformer+LSTM Agent</i> Encoder Decoder	Identical to <i>Transformer Agent</i> . MLP with hidden units [128, 128], followed by LSTM with layer normalization [BKH16] and hidden units [128].
<i>CNN Agent</i> Encoder Decoder	CNN with kernel parameters [(3, 32, 6, 4), (32, 64, 6, 4), (64, 128, 7, 1)] (number of input filters, number of output filters, kernel size, and stride size by ordering). MLP with hidden units [128].
<i>CNN+Transformer Agent</i> Encoder Decoder	CNN with kernel parameters [(3, 32, 4, 4, 0), (32, 64, 4, 4, 0), (64, 128, 3, 2, 1)] (number of input filters, number of output filters, kernel size, stride size, and padding size by ordering); resized to 4×4 slots, concatenated with positional embedding; followed by the encoder of <i>Transformer Agent</i> . Identical to <i>Transformer Agent</i> .
<i>SPACE Agent</i> Encoder Decoder	We adopt the original setup of SPACE [LWP20] for the <code>image_encoder</code> and the <code>what_encoder</code> . We concatenate latent vectors for the shape (Z_{what}) and the presence (Z_{where}) of each object. In sum, there are 8×8 object slots. They are then fed to the encoder of <i>Transformer Agent</i> . Identical to <i>Transformer Agent</i> .

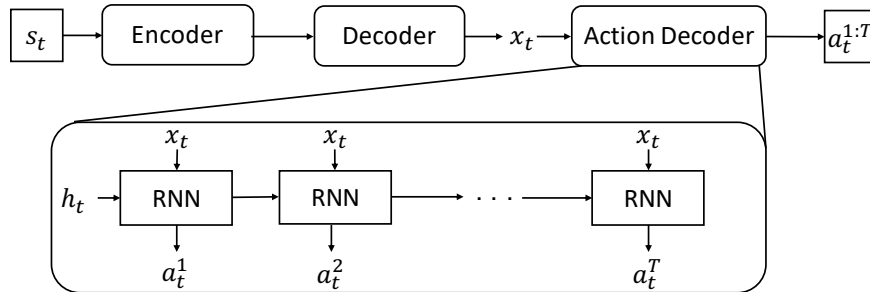


Figure 4.1: Architecture of the actor model, where T is equal to `max_opt_len`.

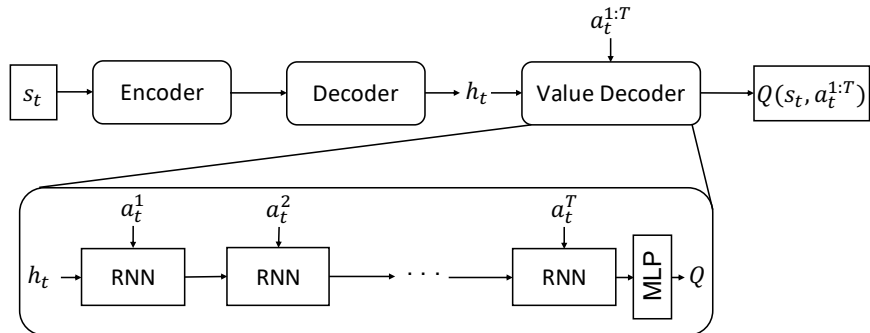


Figure 4.2: Architecture of the critic model, where T is equal to `max_opt_len`.

4.2 Experiment 1: Generalization Tests

By our evaluation protocol, however, these prior models did not demonstrate conclusive evidence to support all three levels of generalization proposed in this paper; hence, the precise level of generalization is obscure. Crucially, neither of them evaluated the learned agents *under limited exposure* to a *complex concept space* as in HALMA.

Table 4.3 shows the full list of agents used in our experiments. All agents are trained with an off-the-shelf reinforcement learning method, TD3 [FHM18]; detailed hyper-parameters defers to 4.2. All agents’ policies converged at the end of training.

To decouple the evaluation of conceptual generalization from perceptual generalization, we first conduct experiments with symbolic one-hot observations, which can be regarded as the ground-truth representation of perception. All agents show relatively high invalid action ratio ρ_a in tests of random split, indicating their understanding of affordance is brittle even with the ground-truth semantics. Under this precondition, we find that all agents can

Table 4.2: Hyper-parameters of TD3

Hyper-parameters	Value
Optimizer	Adam [KB14]
Learning rate for actor	1e-4
Batch size	128
ϵ of Adam	1e-8
Discounting factor	0.95
Initial ϵ for ϵ -greedy	0.1
Ending ϵ for ϵ -greedy	0.95
Decay steps for ϵ -greedy	100,000
Policy update delay	5
Target update rate	0.995
Replay buffer size	10,000

still perform relatively well in terms of goal-reaching ρ_g and efficiency ρ_p in random splits. However, when transferred to our generalization tests, MLP agents exhibits a significant degradation. Agents with LSTM modules, on the contrary, can somehow maintain or even surpass their ρ_g and ρ_p in training problems. One possible explanation to their high ρ_g is: With a memory mechanism, they learn to recover from dead-ends even if they missed the hints at crossings. Even though they also have higher ρ_p than MLP agents, consistent with the findings reported by [WKT16], this measure is still disconcertingly low. Such low performance implies that agents do not understand the concept space well, especially in terms of the temporal grammar. Transformer agents do perform better than MLP agents in generalization tests, but not as good as LSTM agents. In particular, even though [ZRS19] argued that Transformer agents as such may learn to plan, their lower ρ_p in HALMA task implies the opposite, at least under partial observation without a memory mechanism. Combining the benefits from the attention and the memory mechanisms, TRAN+LSTM agents outperform others in almost all generalization tests on both ρ_g and ρ_p . Another interesting phenomenon is: By removing the constraint of *limited exposure* (e.g., we increase the training volume to $10\times$), all agents, no matter what inductive biases are encoded, achieve around 80% measured by ρ_g , and those with LSTM modules have ρ_p at around 45%; see details in Section 4.3.1. Since no state-of-the-art agents could pass the test on ρ_p , we summarize the results of

symbolic experiments as: In the spectrum of model-based vs model-free, emerged strategies still reside on the model-free side of the oracle agent. Significant efforts are needed to devise agents capable of humanlike conceptual and algorithmic generalization.

Under visual observation, however, all agents fail the generalization test when simply connected with a convolutional module, even in the easiest setup (`max_opt_len=1`). Assuming CNNs do not offer sufficient priors to induce an object-oriented, independently disentangled representation, we pretrain a state-of-the-art multi-object segmentation and disentanglement model, SPACE [LWP20], with all visual panels in the training set. The converged model exhibits remarkable generalization in reconstruction, segmentation, and detection, consistent with the results reported by [LWP20]. One would expect that, by connecting the encoder of this powerful pretrained visual module with an RL agent using a Transformer module for the object-oriented encoding, the model would have a superb performance. Counter-intuitively, our results show that SPACE agents perform worse than CNN+TRAN agents even under random split. A further investigation reveals that the latent space of object slots fails to disentangle shapes or colors (*e.g.*, **3** vs **5**), even though they can be substantially distinguished and reconstructed by the strongly nonlinear decoder. This explanation also accounts for SPACE agents’ high invalid action ratio in test problems ($\rho_a = 58.38 \pm 1.20$). In principle, they misunderstand affordance because they fail to recognize “what it is” in the first place. Taking together, we argue that HALMA does extend the evaluation paradigm of perceptual generalization, posing new challenges to the community of unsupervised disentanglement.

Table 4.3: Examples and results of generalization tests (- indicates no problem is dynamically generated)

Test Type & Examples		Models & Results							
		%	SYMBOLIC (max_opt_len=5)				VISUAL (max_opt_len=1)		
			MLP	LSTM	TRAN	TRAN+LSTM	CNN+MLP	CNN+TRAN	SPACE
T	Training problems	ρ_a ↓	5.22±4.11	12.12±2.14	14.57±6.77	13.05±3.09	14.39±7.22	10.29±2.61	16.45±2.65
		ρ_g ↑	99.23±0.63	57.22±3.07	93.85±1.26	72.33±5.79	75.76±4.77	58.33±4.19	16.33±0.94
		ρ_p ↑	71.67±1.73	50.91±3.54	67.89±0.63	63.97±5.84	63.77±2.68	35.31±3.00	12.02±1.17
RT	Random split	ρ_a ↓	37.02±1.52	23.91±2.10	34.85±4.45	37.69±2.90	86.70±2.30	56.91±7.92	58.38±1.20
		ρ_g ↑	51.00±2.21	57.78±3.49	82.82±0.96	54.00±2.94	7.58±0.43	14.00±4.24	3.67±0.47
		ρ_p ↑	54.91±2.85	45.15±1.46	58.07±1.01	40.13±2.52	5.09±1.17	8.33±1.96	2.66±0.19
ST	$\langle 3, 5, 1, 7 \rangle \in \text{MEM}$, test $\langle 3, 5, 4, 2 \rangle \notin \text{MEM}$. $3 < 5 \in \text{KB}_S$, test $\langle 3, 5 \rangle \notin \text{KB}_S$.	ρ_g ↑	55.00±7.07	50.00±8.16	41.67±8.50	66.67±13.12	0.00±0.00	0.00±0.00	0.00±0.00
		ρ_p ↑	19.90±2.18	24.02±7.20	16.34±3.90	35.74±5.85	0.00±0.00	0.00±0.00	0.00±0.00
		ρ_g ↑	25.00±8.16	63.33±6.24	43.33±6.23	78.33±2.36	0.00±0.00	0.00±0.00	0.00±0.00
AFT	$5 = 3 + \blacktriangle : \square \in \text{KB}_{Avc}$, test $\langle 3, 5 \rangle \notin \text{KB}_S$. $\{5 = \blacktriangle : \square + \square, 3 = \blacktriangle : \square\}$ $\subset \text{KB}_{Avc}$, test $\langle 3, 5 \rangle \notin \text{KB}_S$. $5 = 3 + \blacktriangle : \square \in \text{KB}_{Avc}$, test $\langle 3, 5 \rangle \notin \text{KB}_S$. $\{5 = \blacktriangleright : \square + \square, 3 = \blacktriangleright : \square\}$ $\subset \text{KB}_{Avc}$, test $\langle 3, 5 \rangle \notin \text{KB}_S$.	ρ_g ↑	41.67±2.36	60.00±10.80	36.67±8.50	58.33±10.27	0.00±0.00	0.00±0.00	0.00±0.00
		ρ_p ↑	15.10±0.35	28.91±7.62	14.01±3.75	27.11±2.12	0.00±0.00	0.00±0.00	0.00±0.00
		ρ_g ↑	31.67±8.50	45.00±10.80	43.33±6.24	71.67±6.24	0.00±0.00	0.00±0.00	0.00±0.00
AnT	$3 < 4, 4 < 5, 3 < 5, 6$ $< 7, 7 < 8 \rangle \subset \text{KB}_{STvAft}$, test $\langle 6, 8 \rangle \notin \text{KB}_{STvAft}$.	ρ_g ↑	11.68±3.34	17.15±5.82	17.86±3.02	35.40±3.71	0.00±0.00	0.00±0.00	0.00±0.00
		ρ_g ↑	6.67±2.36	100.00±0.00	25.00±0.00	-	0.00±0.00	0.00±0.00	0.00±0.00
		ρ_p ↑	1.48±0.52	51.86±0.18	5.83±0.24	-	0.00±0.00	0.00±0.00	0.00±0.00
AnT	$\{5 = \blacktriangleright : \square + \square, 3 = \blacktriangleright : \square\}$ $\subset \text{KB}_{Avc}$, test $\langle 3, 5 \rangle \notin \text{KB}_S$. $\{3 < 4, 4 < 5, 3 < 5, 6$ $< 7, 7 < 8 \rangle \subset \text{KB}_{STvAft}$, test $\langle 6, 8 \rangle \notin \text{KB}_{STvAft}$.	ρ_g ↑	0.00±0.00	86.67±9.43	50.00±0.00	-	0.00±0.00	0.00±0.00	0.00±0.00
		ρ_p ↑	0.00±0.00	29.89±2.18	10.00±0.00	-	0.00±0.00	0.00±0.00	0.00±0.00
		ρ_g ↑	35.00±7.07	48.33±4.71	41.67±2.36	41.67±13.12	0.00±0.00	-	0.00±0.00
AnT	$\{3 < 4, 4 < 5, 3 < 5, 6$ $< 7, 7 < 8 \rangle \subset \text{KB}_{STvAft}$, test $\langle 6, 8 \rangle \notin \text{KB}_{STvAft}$.	ρ_p ↑	12.19±1.84	21.84±0.53	14.45±1.66	22.03±6.64	0.00±0.00	-	0.00±0.00
		ρ_g ↑	12.19±1.84	21.84±0.53	14.45±1.66	22.03±6.64	0.00±0.00	-	0.00±0.00
		ρ_p ↑	12.19±1.84	21.84±0.53	14.45±1.66	22.03±6.64	0.00±0.00	-	0.00±0.00

4.3 Experiment 2: Ablation Studies

4.3.1 Ablation on the Volume of Training Set

The thesis argument of our work is that humanlike agents shall generalize their understanding under limited exposure to the underlying concept spaces. To further investigate how the degree of exposure would affect agents performance in HALMA, we first conduct an ablations study with different numbers of training mazes. Specifically, we experiment with four setups of the maze quantity for agents to explore during training: 100, 300, 500, 1000 (results of 100 training mazes are reused from the main experiment as it is our default setting). Here we only evaluate agents with symbolic input: MLP agents, LSTM agents, Transformer agents and Transformer+LSTM agents. We report the three measures ρ_a , ρ_g and ρ_p with all the testing protocols (training problems, problems from random split in the problem space and dynamically-generated testing problems) in Fig. 4.3. Note that measures in dynamically-generated tests are merged across subtests for better comparison.

The results read that, all agents could gain a performance boost with increased exposure

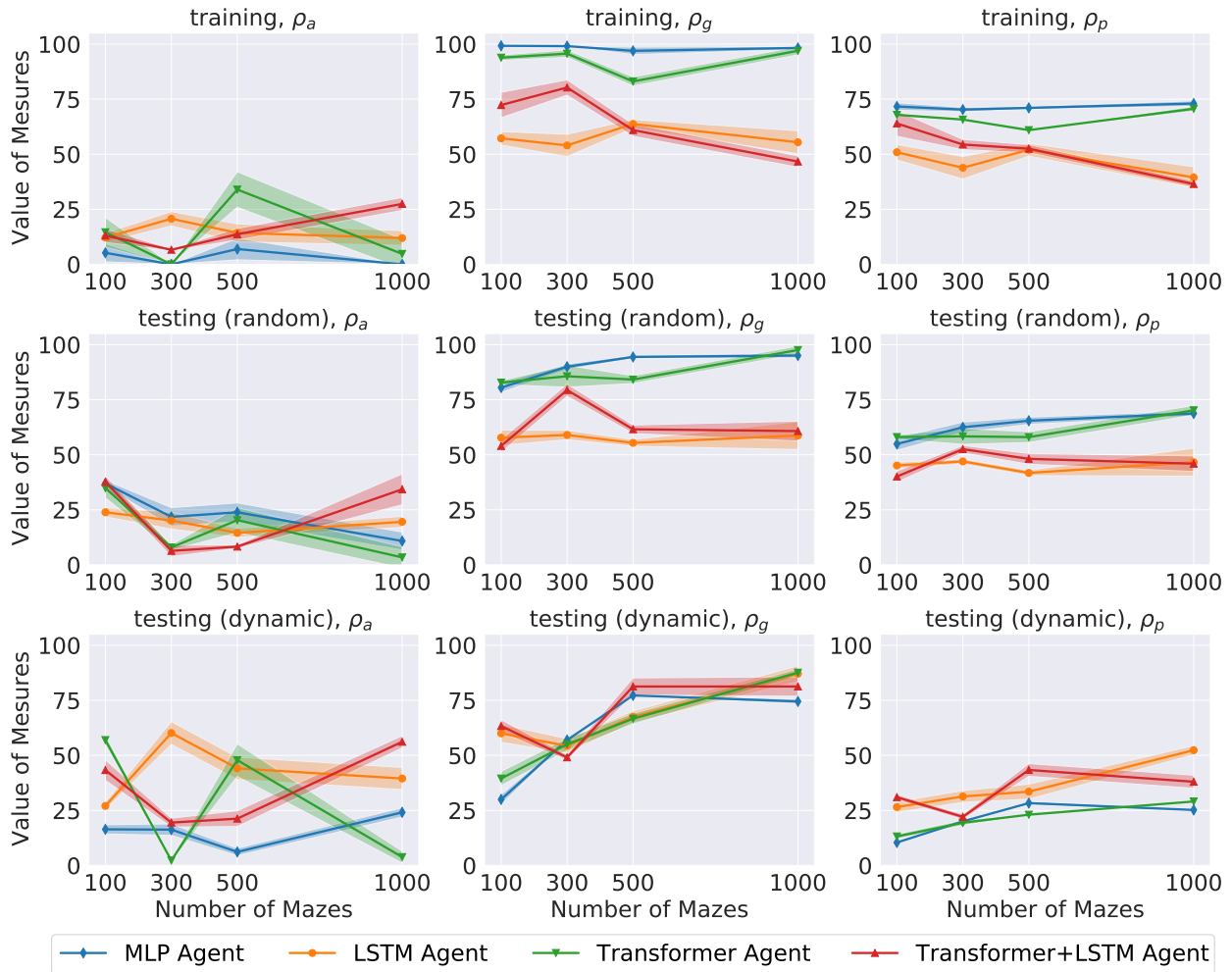


Figure 4.3: Ablation study of different number of training mazes.

during training. Specifically, there is a significant promotion for the metric of goal reaching rate ρ_g in the challenging dynamic testing (from 30-60% to 80%). More interestingly, starting from 300 training mazes, the distinction between different inductive biases vanishes. While the efficiency ratio ρ_p could also benefit from increased exposure, it reaches only around 50% at best. As for the ratio of invalid moves ρ_a , even though it reaches around 10% in random split for stateless agent when trained with 1000 mazes, no clear trend can be detected in dynamic testing overall, which may suggest agents' limitation in understanding affordance with the temporal grammar or under the long-tail distribution of digits.

4.3.2 Ablation on the Maximum Option Length

Our design to include the notion of option challenges agents’ understanding in the temporal grammar and the causal structure. To further illustrate the difficulty of this specific challenge, we also perform an ablation study on three setups of maximum option length `max_opt_len`. In general, agents’ performance degrades on all metrics with `max_opt_len` increases. In particular, the ratio of invalid moves ρ_a increases and the efficiency ratio ρ_p drops significantly since `max_opt_len=3` in dynamic testing, suggesting that agents all have hard time understanding either the temporal grammar or the causal structure of HALMA. These results validate our argument that significant efforts are still in need for humanlike abstraction learning. Therefore, we choose to make the length of 5 as our default setting in the main paper so as to make HALMA a more challenging territory.

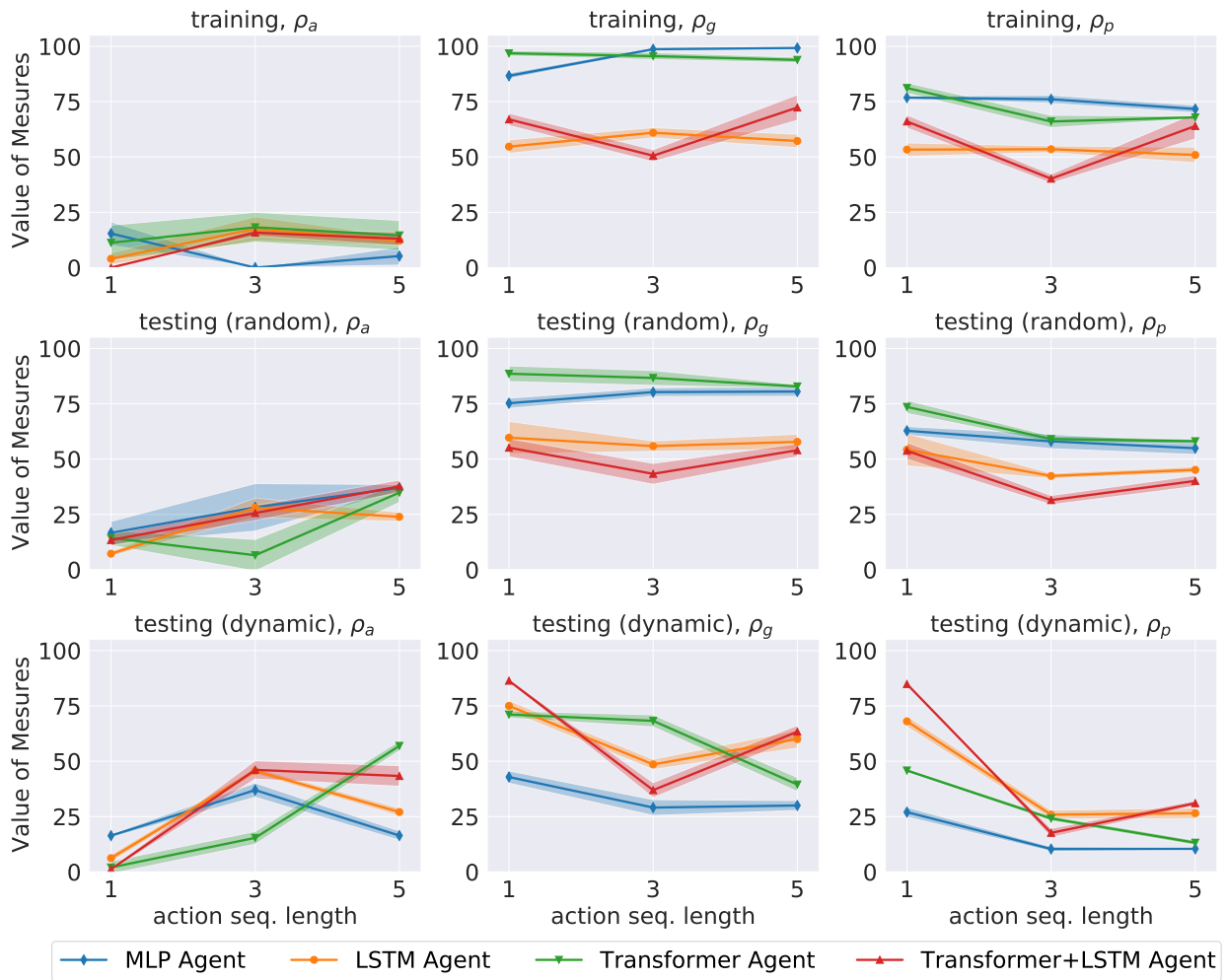


Figure 4.4: Ablation study of different `max_opt_len` (symbolic observations).

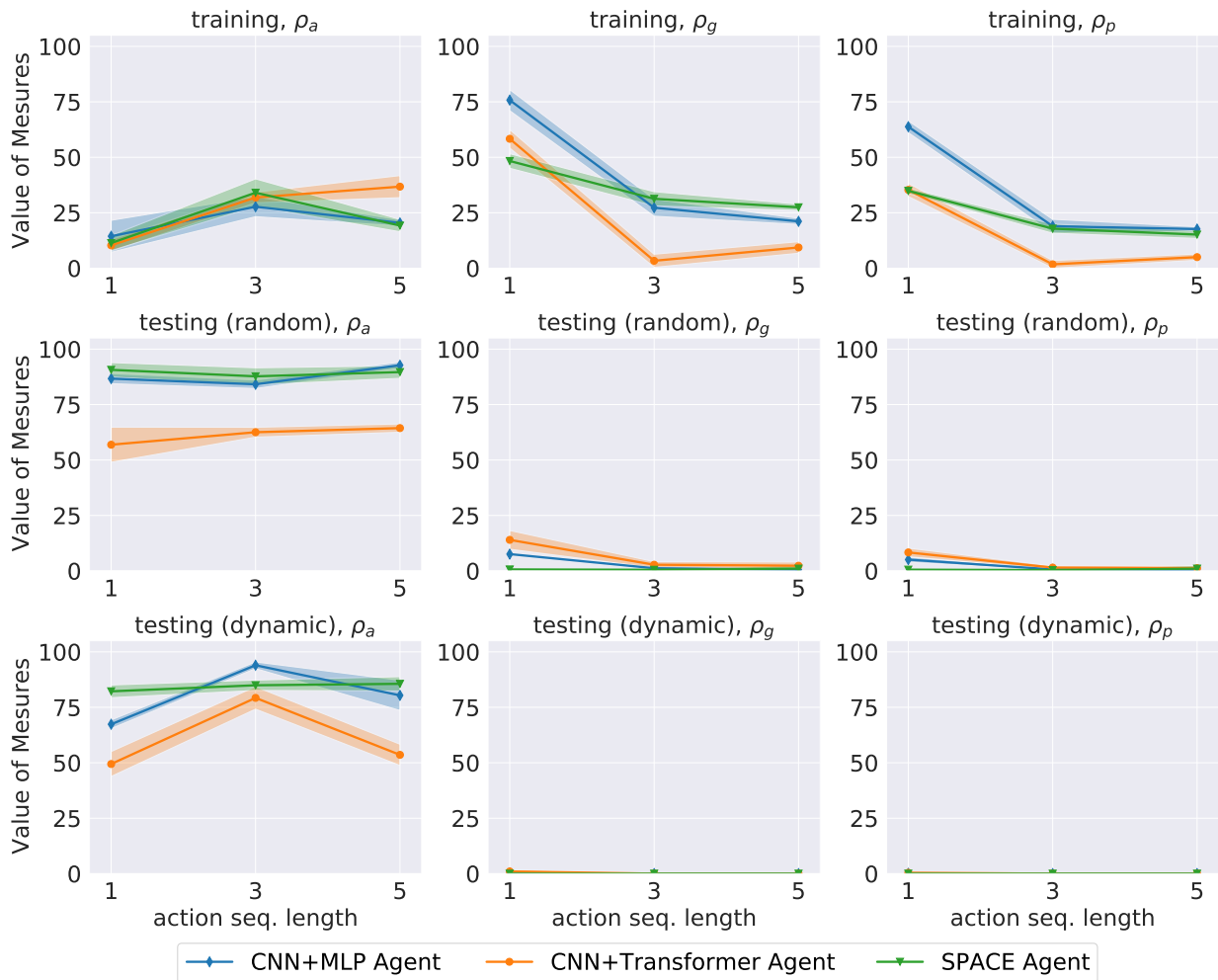


Figure 4.5: Ablation study of different `max_opt_len` (visual observations).

CHAPTER 5

Conclusive Remarks

In spite of its synthetic nature, we believe HALMA is an impeccable testbed for rapid problem solving that resembles real-world ones. The dedicated design of its internal state facilitates in-depth and comprehensive analyses on agents' capacity in concept development, abstract reasoning, and meta learning that are otherwise impossible with existing problem-solving tasks. Agents can only pass the dynamically generated generalization tests if they possess adequate capacity to *understand* the abstract structure of this task and build a powerful solver upon this understanding. Our experiments demonstrate the inefficacy of model-free reinforcement learning agents in generalizing their understanding, even when incorporated with generic inductive biases. Towards this end, we would like to invite colleagues across the machine learning community to join our challenge.

REFERENCES

- [AGP15] Aurore Avarguès-Weber, Martin Giurfa, Joshua Plotnik, Nicola S Clayton, Robert Seyfarth, Dorothy L Cheney, Brad Mahon, H Clark Barrett, Pascal Boyer, Jerry A Fodor, et al. *The conceptual mind: New directions in the study of concepts*. MIT Press, 2015.
- [Ano21] Anonymous. “{HALMA}: Humanlike Abstraction Learning Meets Affordance in Rapid Problem Solving.” In *Submitted to International Conference on Learning Representations*, 2021. under review.
- [BC14] Matthew M Botvinick and Jonathan D Cohen. “The computational and neural basis of cognitive control: charted territory and new frontiers.” *Cognitive Science*, **38**(6):1249–1285, 2014.
- [BHB18] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. “Relational inductive biases, deep learning, and graph networks.” *arXiv preprint arXiv:1806.01261*, 2018.
- [BHS18] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. “Measuring abstract reasoning in neural networks.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2018.
- [BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization.” *arXiv preprint arXiv:1607.06450*, 2016.
- [BMN19] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. “Systematic Generalization: What Is Required and Can It Be Learned?” In *International Conference on Learning Representations*, 2019.
- [BMW19] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. “Monet: Unsupervised scene decomposition and representation.” *arXiv preprint arXiv:1901.11390*, 2019.
- [Car09] Susan Carey. *The origin of concepts*. Oxford Press, 2009.
- [CGC14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. “Empirical evaluation of gated recurrent neural networks on sequence modeling.” *arXiv preprint arXiv:1412.3555*, 2014.
- [Cho86] Noam Chomsky. *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group, 1986.

- [CKH19] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. “Quantifying generalization in reinforcement learning.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2019.
- [CLS20] Devendra Singh Chaplot, Lisa Lee, Ruslan Salakhutdinov, Devi Parikh, and Dhruv Batra. “Embodied Multimodal Multitask Learning.” In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 2020.
- [Deh11] Stanislas Dehaene. *The number sense: How the mind creates mathematics*. OUP USA, 2011.
- [EHW16] SM Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. “Attend, infer, repeat: Fast scene understanding with generative models.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [Elm90] Jeffrey L Elman. “Finding structure in time.” *Cognitive Science*, **14**(2):179–211, 1990.
- [FC03] Lisa Feigenson and Susan Carey. “Tracking individuals via object-files: evidence from infants’ manual search.” *Developmental Science*, **6**(5):568–584, 2003.
- [FHM18] Scott Fujimoto, Herke Hoof, and David Meger. “Addressing Function Approximation Error in Actor-Critic Methods.” In *International Conference on Machine Learning*, pp. 1587–1596, 2018.
- [Fla63] John H Flavell. *The developmental psychology of Jean Piaget*. D Van Nostrand, 1963.
- [FP88] Jerry A Fodor, Zenon W Pylyshyn, et al. “Connectionism and cognitive architecture: A critical analysis.” *Cognition*, **28**(1-2):3–71, 1988.
- [FPP11] Norm Ferns, Prakash Panangaden, and Doina Precup. “Bisimulation metrics for continuous Markov decision processes.” *SIAM Journal on Computing*, **40**(6):1662–1714, 2011.
- [GB20] Anirudh Goyal and Yoshua Bengio. “Inductive Biases for Deep Learning of Higher-Level Cognition.” *arXiv preprint arXiv:2011.15091*, 2020.
- [GFP19] Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. “How efficiency shapes human language.” *Trends in cognitive sciences*, **23**(5):389–407, 2019.

- [GKB19] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. “DeepMDP: Learning Continuous Latent Space Models for Representation Learning.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2019.
- [GKK19] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. “Multi-Object Representation Learning with Iterative Variational Inference.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2019.
- [GMG19] Arthur Guez, Mehdi Mirza, Karol Gregor, Rishabh Kabra, Sebastien Racaniere, Theophane Weber, David Raposo, Adam Santoro, Laurent Orseau, Tom Eccles, et al. “An Investigation of Model-Free Planning.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2019.
- [Gre93] Ulf Grenander. *General pattern theory: A mathematical study of regular structures Oxford mathematical monographs*. Oxford University Press: Clarendon, 1993.
- [Gri20] Thomas L Griffiths. “Understanding Human Intelligence through Human Limitations.” *Trends in Cognitive Sciences*, 2020.
- [GTF08] Noah D Goodman, Joshua B Tenenbaum, Jacob Feldman, and Thomas L Griffiths. “A rational analysis of rule-based concept learning.” *Cognitive Science*, **32**(1):108–154, 2008.
- [HMP17] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. “beta-VAE: Learning basic visual concepts with a constrained variational framework.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2017.
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory.” *Neural Computation*, **9**(8):1735–1780, 1997.
- [JHM17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [KA13] Kari S Kretch and Karen E Adolph. “Cliff or step? Posture-specific learning at the edge of a drop-off.” *Child Development*, **84**(1):226–240, 2013.
- [KAC20] Khimya Khetarpal, Zafarali Ahmed, Gheorghe Comanici, David Abel, and Doina Precup. “What can I do here? A Theory of Affordances in Reinforcement Learning.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2020.

- [Kar94] By A Karmiloff-Smith. “Beyond modularity: A developmental perspective on cognitive science.” *European journal of disorders of communication*, **29**(1):95–105, 1994.
- [KB14] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization.” *arXiv preprint arXiv:1412.6980*, 2014.
- [KLC98] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. “Planning and acting in partially observable stochastic domains.” *Artificial Intelligence*, **101**(1-2):99–134, 1998.
- [KSM17] Ken Kansky, Tom Silver, David A Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, and Dileep George. “Schema networks: zero-shot transfer with a generative causal model of intuitive physics.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2017.
- [KTG92] Daniel Kahneman, Anne Treisman, and Brian J Gibbs. “The reviewing of object files: Object-specific integration of information.” *Cognitive Psychology*, **24**(2):175–219, 1992.
- [LB18] Brenden Lake and Marco Baroni. “Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2018.
- [LBL19] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. “Challenging common assumptions in the unsupervised learning of disentangled representations.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2019.
- [LHT18] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. “Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input.” In *International Conference on Learning Representations (ICLR)*, 2018.
- [LST15] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. “Human-level concept learning through probabilistic program induction.” *Science*, **350**(6266):1332–1338, 2015.
- [LUT17] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. “Building machines that learn and think like people.” *Behavioral and Brain Sciences*, **40**, 2017.
- [LWL06] Lihong Li, Thomas J Walsh, and Michael L Littman. “Towards a Unified Theory of State Abstraction for MDPs.” In *International Symposium on Artificial Intelligence and Mathematics*, 2006.

- [LWP20] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. “SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition.” In *International Conference on Learning Representations (ICLR)*, 2020.
- [MA18] Igor Mordatch and Pieter Abbeel. “Emergence of Grounded Compositional Language in Multi-Agent Populations.” In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- [Mar82] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., USA, 1982.
- [Mar18] Gary F Marcus. *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press, 2018.
- [MBN10] James L McClelland, Matthew M Botvinick, David C Noelle, David C Plaut, Timothy T Rogers, Mark S Seidenberg, and Linda B Smith. “Letting structure emerge: connectionist and dynamical systems approaches to cognition.” *Trends in cognitive sciences*, **14**(8):348–356, 2010.
- [MHR20] James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. “Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models.” *Proceedings of the National Academy of Sciences*, **117**(42):25966–25974, 2020.
- [MLB08] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. “Learning object affordances: from sensory–motor coordination to imitation.” *Transactions on Robotics (T-RO)*, **24**(1):15–26, 2008.
- [MPV17] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andy Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. “Learning to Navigate in Complex Environments.” In *International Conference on Learning Representations (ICLR)*, 2017.
- [Qui60] Willard Van Orman Quine. *Word and object*. MIT press, 1960.
- [RNB05] Nicolas P Rougier, David C Noelle, Todd S Braver, Jonathan D Cohen, and Randall C O’Reilly. “Prefrontal cortex and flexible cognitive control: Rules without symbols.” *Proceedings of the National Academy of Sciences (PNAS)*, **102**(20):7338–7343, 2005.
- [RWK18] Samuel Ritter, Jane Wang, Zeb Kurth-Nelson, Siddhant Jayakumar, Charles Blundell, Razvan Pascanu, and Matthew Botvinick. “Been There, Done That: Meta-Learning with Episodic Recall.” In *Proceedings of International Conference on Machine Learning (ICML)*, 2018.

- [San08] Scott Patrick Sanner. *First-order decision-theoretic planning in structured relational environments*. PhD thesis, University of Toronto, 2008.
- [SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [SKM19] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. “Habitat: A platform for embodied ai research.” In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9339–9347, 2019.
- [SPS99] Richard S Sutton, Doina Precup, and Satinder Singh. “Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning.” *Artificial intelligence*, **112**(1-2):181–211, 1999.
- [SZW96] Juergen Schmidhuber, Jieyu Zhao, and MA Wiering. “Simple principles of meta-learning.” Technical report, IDSIA, 1996.
- [TWT16] Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, and Pieter Abbeel. “Value iteration networks.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [VEB17] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. “Starcraft ii: A new challenge for reinforcement learning.” *arXiv preprint arXiv:1708.04782*, 2017.
- [VSP17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is all you need.” In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [WHA18] Greg Wayne, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwinska, Jack Rae, Piotr Mirowski, Joel Z Leibo, Adam Santoro, et al. “Unsupervised predictive memory in a goal-directed agent.” *arXiv preprint arXiv:1803.10760*, 2018.
- [WKK18] Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. “Prefrontal cortex as a meta-reinforcement learning system.” *Nature Neuroscience*, **21**(6):860–868, 2018.
- [WKM20] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. “DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames.” In *International Conference on Learning Representations*, 2020.

- [WKT16] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. “Learning to reinforcement learn.” *arXiv preprint arXiv:1611.05763*, 2016.
- [YWF19] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. “Visual Semantic Navigation using Scene Priors.” In *International Conference on Learning Representations*, 2019.
- [ZGF20] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, Josh B Tenenbaum, and Song-Chun Zhu. “Dark, Beyond Deep: A Paradigm Shift to Cognitive AI with Human-like Common Sense.” *Engineering*, **6**(3):310–345, 2020.
- [ZGJ19] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. “Raven: A dataset for relational and analogical visual reasoning.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [ZMC20] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarín Gal, and Sergey Levine. “Learning Invariant Representations for Reinforcement Learning without Reconstruction.” *arXiv preprint arXiv:2006.10742*, 2020.
- [ZRS19] Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. “Deep reinforcement learning with relational inductive biases.” In *International Conference on Learning Representations (ICLR)*, 2019.
- [ZZZ15] Yixin Zhu, Yibiao Zhao, and Song-Chun Zhu. “Understanding tools: Task-oriented object modeling, learning and recognition.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.