Relational Concepts in Deep Reinforcement Learning: Emergence and Representation

Thomas Schnürer[1,2], Malte Probst[2], and Horst Michael Groß[1]

[1] Neuroinformatics and Cognitive Robotics Lab, Ilmenau University of Technology, Ilmenau, Germany
{thomas.schnuerer, horst-michael.gross}@tu-ilmenau.de.de
[2] Honda Research Institute Europe GmbH, 63073 Offenbach/Main, Germany
malte.probst@honda-ri.de

1 Introduction

Current AI systems are usually built to solve a single, well-specified task or need a carefully controlled environment. Even though impressive results have been achieved for such tasks lately, the necessary knowledge is usually learned from scratch, bound to a specific task & environment, and requires a lot of resources for training. For an intelligent agent in a complex environment facing a variety of tasks, such approaches entail a number of severe limitations, including generality, reusability, and interpretability.

In an effort to address some of these limitations, we want to move towards a separation of general knowledge and task specific knowledge. General knowledge is grounded in the agent's experience and represented in a distributed, sub-symbolic way. Similar to how humans are able to quickly adapt to new problems [1], task specific knowledge is then learned "on top" by leveraging these pre-existing general concepts. The intuition is that this transfer learning approach will allow for simpler, better understandable task representations and consequently increase the speed and robustness of their learning process[2].

In this setting, we regard concepts to be one possible type of such general knowledge. While the problem of acquiring grounded concepts is traditionally seen in a top-down manner [3], we follow recent advances that instead suggest a shift towards bottom-up approaches [4]. With this in mind, reinforcement learning becomes a natural choice as a training method. There, an agent is interacting with its environment in an effort to optimize a policy. However, instead of training labels (e.g. the correct concepts or actions for a specific situation), the agent only receives a reinforcement signal (reward or punishment) after each step. Because of that, very little prior knowledge (e.g. human concepts) is imposed top-down and the agent has to learn its own concepts for understanding the environment. Furthermore, the reinforcement signal could even be exchanged for intrinsic motivation [5] later on.

As a sub-type of such general concepts, we focus on relational concepts that exist between two or more objects. Such relations could be "unlock" (key →lock), "container" (books →box), or "tool" (screwdriver →screw). While specific relations between instances are usually task dependent (e.g. repair laptop: screwdriver →screw; clean desk: screwdriver →drawer), we aim for general relational concepts that are not bound to specific instances or tasks (e.g. "inside", "bigger", "same type")

Importantly, since the agent itself can also be seen as an object with both extrinsic and intrinsic properties, relations between the agent and objects can encode affordances [6] (e.g. "open", "move", "pick up") and other higher level concepts. Therefore, we see relational concepts between objects as a first step towards a mechanism for more general relational concepts between perceptions, objects, situations, memories, goals, skills, actions, etc.

2 State of the Art

Approaches leveraging relational architectures have recently shown to solve tasks of impressing complexity [7][8]. While this success suggests that such approaches must have learned a number of complex concepts "on their own", they are limited to a single task and environment. Due to their design, potentially learned concepts are entangled with perception, action policies, and also with each other. This makes it very difficult to reuse certain parts across tasks or to see what kinds of concepts they have learned and how they are represented. Our goal is to assess how well such architectures are suited for transfer learning while specifically focusing on transparency and interpretability.

3 Experiments

In order to create some understanding for the inner workings, capabilities, and differences of relational architectures, we employ a minimalistic environment that allows us to zero in on specific aspects of such architectures. With this, we investigate the representations that are formed as well as the conditions that lead to them. Finally, we use this knowledge to explore potential inductive biases (e.g. purposeful architecture modifications) that can help architectures to form high-level relational concepts instead of instance relations. For example, we are considering different ways of forming the neuronal connections for transforming observations into states. Our experiments will follow the guidelines of the following research hypotheses:

1. Current reinforcement learning architectures tend to focus more on task specific instance relations rather than relational concepts
2. It is possible to train an architecture with reinforcement learning in an exhaustive way that enforces bottom up general relational concepts
3. It is possible to introduce an inductive bias to such an architecture that allows for relational concepts to emerge without the need of 2. and provides higher interpretability

Our minimalistic environment is derived from BoxWorld [8] and contains a number of colored keys and locks. By default, a key can open a lock if both have the same color (e.g. red key →red lock). Since an environment can contain multiple keys and locks, the agent must navigate it and open each lock in order to solve the task. To investigate whether an architecture learns instance relations (e.g. "red key" →"red lock") or the underlying relational concept (keys and locks match if they have the same color), we introduce a second case. Here, locks and keys do not match by color but are randomly assigned at the beginning of the training. In this case, all pairs of keys and locks have to be memorized since there is no higher concept to learn. While both cases are virtually identical for architectures focusing on instance relations, architectures that are able to learn higher concepts are expected to perform significantly better in the first case (matching colors). Based on this setup we derive a number of experiments to examine current architectures and explore possible inductive biases. Across all experiments and architectures, we compare the performance on unseen key-lock-pairs as well as the representations that are formed by these architectures.

Our intermediate results indicate that our first hypothesis is correct and architectures like the Transformer [8] or Relation Network [7] do indeed focus on instance relations and have difficulties learning higher level concepts. We are following up on these indications with further experiments also regarding research hypotheses two and three.

4 Conclusion

At the workshop, we will present the ongoing empirical results of our experiments as well as our gained insights. More specifically, we want to highlight the capabilities and limitations of current ML architectures for learning and representing general relational concepts in a bottom up fashion. For this, we analyze the types and qualities of representations that different architectures foster in regards to the concepts they can or do encode. Additionally, we gather empiric data on the training process that is able to allow or prevent the emergence of these concepts. We are looking forward to contributing our expertise as well as to the feedback and interesting discussions that arise from the workshop.

References

1. Dubey et al.: Investigating Human Priors for Playing Video Games. ICML (2018).
2. Mao et al.: Universal Agent for Disentangling Environments and Tasks. ICLR (2018)
3. Steels et al.: The Symbol Grounding Problem Has Been Solved, so Whats Next?. In: Symbols and Embodiment: Debates on Meaning and Cognition (2019).
4. Taniguchi et al.: Symbol Emergence in Cognitive Developmental Systems: a Survey. In: IEEE TCDS (2018)
5. Burda et al.: Large-Scale Study of Curiosity-Driven Learning. (2018)
6. Zech et al.: Computational Models of Affordance in Robotics: A Taxonomy and Systematic Classification. In: Adaptive Behavior **25**(5): 235–71 (2017)
7. Santoro et al.: A Simple Neural Network Module for Relational Reasoning. In: Advances in Neural Information Processing Systems: 4968–77 (2019)
8. Zambaldi et al.: Deep Reinforcement Learning with Relational Inductive Biases. ICLR (2019)