Ph.D. Proposal

Title: Semantic Representations for Interpretable Reinforcement Learning

Supervisor (hdr): Pr. Philippe Preux
Co-supervisor: Dr. Riad Akrour

Nov. 27th, 2021

Keywords: Reinforcement learning, Artificial Intelligence, policy explanability, hybrid (symbolic/continuous) learning

Contact information:

- Supervisor: Philippe Preux' email address: `philippe.preux@univ-lille.fr` and website: `https://philippe-reux.github.io`
- Co-supervisor: Riad Akrour's email address: `riad.akrour@inria.fr` and website: `https://www.ias.informatik.tu-darmstadt.de/Team/RiadAkrour`

- The student will be employed by the Université de Lille, France
- The student will be part of the École Doctorale MADIS
- The Ph.D. will be made in the research group Scool: Scool website: `https://team.inria.fr/scool`
- The student will be a member of both CRIStAL and Inria-Lille labs.

Abstract: Automating decision making is an appealing prospect of AI. However, the solution produced by an AI system may need to be inspected by a human expert before its deployment in the real world. To this end, the AI needs to return a policy in a format semantically meaningful to the human. In addition, the complexity of the policy needs to be controlled to take into account the bounded capacities of humans. In the machine learning and reinforcement learning literature, there has been an increasing amount of research for learning disentangled and semantically meaningful state representations, and for learning compact and human readable policies. In this Ph.D. proposal we aim at i) investigating graph representations of the world to identify its key components and relations, forming the basis of states over which interpretable policies are defined. ii) Extending the notion of disentanglement to hierarchical RL policies, forming the basis of actions over which interpretable policies are defined. iii) Learning compact and human readable policies by RL in game domains with image inputs, which is currently only achievable

by non-interpretable RL methods. The scope of the research is predominantly algorithmic and will be evaluated on game domains, but we discuss the potential impact of our proposal on robotics research and applications such as agriculture and health, in which our team is invested through various on-going projects.

1 Context and related work

Intelligent agents will assist, sometimes, replace humans on a growing number of tasks such as drug discovery or crop harvesting. The challenge of these applications is the diversity of scenarios they present. Diversity that a computer program can only cope with if it is able to self-improve from data. Reinforcement learning (RL) is one such framework capable of learning sequential decision making policies by interacting with an environment. RL has had several achievements, notably in game domains. However, a wider application of RL to real world problems is held back in many cases by: (P1) the reliance of RL on black-box models that are hard to interpret. Whereas in some applications, it might be desirable for the AI to return a policy in a human readable format to be able to understand it, predict its behavior or assess its safety. (P2) The performance of these models collapses when changes, considered insignificant for humans, are introduced in the environment. As such, even if current algorithms are able to self-improve from data, they often learn overly specialized models that do not generalize to slightly different datasets, as demonstrated for example in game domains [4]. We propose to study these two specific limitations jointly because we think they have a common solution: to learn a semantic representation of the AI's environment. This semantic representation will allow to i) express the learned policy in an abstract and compact manner more amenable to human inspection, ii) disentangle independent components of the environment to obtain representations robust to partial changes in the data. Interpretable RL and disentangled representations are active research areas that are briefly summarized in the following.

Interpretable RL. There are two main ways of achieving interpretability in RL: i) learning a black-box policy (e.g. a neural network) and making it interpretable post-hoc [10, 3, 7], by imitation learning of a more interpretable structure (e.g. a small decision tree) or ii) by learning an interpretable policy from scratch [6, 2]. We first note that in both lines of research, one of the biggest technical challenge of interpretable RL is the omnipresence of discrete components in the parameterization of interpretable policies, such that one would be hard pressed to find a work that does not use a discrete optimizer on top of a gradient descent one to learn these policies. Secondly, interpretability of the policy is often associated to a constraint on the representational power of the underlying model. As such, there is an inherent trade-off between the performance of the policy and its interpretability. However, the optimal policy in a restricted policy class might need to follow a vastly different strategy than that of a richer class. Finding these different strategies can only be achieved using RL instead of imitation learning. Unfortunately, learning an interpretable policy with RL is significantly more challenging (need to interweave discrete optimization with RL that is sufficiently challenging on its own) and was not demonstrated on environments with complex inputs such as images.

Disentangled representations. They aim at learning latent representations of the data that have a factorized distribution [5]. The factorized nature encodes the intuitive notion that the true generative model of a given data type, e.g. images, is a product of independent factors such as object shape, position, texture, viewpoint and so on. Despite the popularity of disentanglement as a research field, it is argued in [8] that the problem itself is ill-posed, save for when domain specific assumptions are made on the data types and models. In this thesis, we will focus on multi-object images generated by computer games, that are also widespread RL benchmarks. Specifically, the representation learning part will leverage recent models [9, 11] that disentangle the different objects in a game image, and for each object disentangles its spatial information and color from its shape. We are also currently investigating similar models in the context of image-based RL benchmarks. These disentangled factors of the representation have a clear meaning for humans and are as such relevant for interpretable RL. Another important property of these representations is that shapes are encoded using a fixed size codebook through a clustering method. Uniquely identifying objects in a self-supervised way will serve as a perfect basis for extending their semantic representation with RL specific properties.

2 Research goals and impact

There are three main research goals in the project. i) Extend the existing object-centric semantic representations to incorporate additional properties meaningful to a sequential decision making task. ii) Extend the notion of disentanglement to (hierarchical) RL by learning sub-policies that can independently alter one factor of variation at a time. iii) Learn compact human readable policies by RL in game domains with image inputs.

G1) Key component identification in RL. Leveraging previously mentioned work in disentangled representations, we are able to uniquely identify objects in an image. These objects will be seen as vertices in a graph. In the context of RL, we will be interested in learning two types of dynamic graphs[1]: one modeling the transition function and one modeling the reward function. Edges in these graphs express functional dependencies. For example, the transition function's dynamic graph in the Pong Atari game will have an edge between the ball and the player's pad before they collide to express that the ball's next state will depend on its current state and the state of the pad—and nothing else. Modelling the transition and reward functions is typically useful in the context of model-based RL but this research topic is orthogonal to this proposal. The use of these graphs in the context of interpretable RL is twofold. First, it allows to filter out components that do not impact future rewards (either directly or indirectly). For instance the score in Pong is indicative of past rewards but influences neither the immediate future rewards nor the state of elements that do impact future rewards. Secondly, it helps the identification of components that interact with each others and under which conditions they do so. Identifying these key components and key relations between them will greatly reduce the (typically combinatorial) search space of interpretable policies. Their discrete optimization becomes suddenly tractable for some tasks.

G2) Disentanglement in the state-action space. Current interpretable policy structures provide a transparent relation between states and actions—expressed by decision trees, symbolic mathematical equations or proximity to a set of prototypical situations—but this might not be enough to understand the strategy followed by the AI. One workaround is for these policies to produce more abstract actions, in the form of sub-goals, that better explain the intent of the AI at a given state. Autonomously finding relevant sub-goals has been a long lasting research topic in hierarchical RL. We propose to leverage the learned disentangled representation of the state to learn 'disentangled sub-policies': sub-policies that are able to change a given disentangled feature of a state while minimally changing the rest. The main interest of these sub-policies is that because they only affect one part of the representation, they can easily be chained to reach any target representation that requires alteration of multiple parts of the current representation. Of course, sometimes the absolute value of a feature (e.g. position of a ball in Pong) is less important than its relative value w.r.t. a set of other features (e.g. the position of the ball w.r.t. the position of the player's pad). Thus, more generally we want sub-policies that are able to bring a subset of features to a target sub-goal while minimally altering the rest of the state representation. By having to consider

[1] Dynamic graph here means that the edges between vertices may change at every time-step.

combinations of features, the amount of sub-goals can become prohibitively large but we can leverage the set of key components and key relations as defined in G1 to prioritize these sub-problems.

G3) Learning compact interpretable policies from visual states. Having decomposed the state into a set of independent factors (G1), and having discovered a set of sub-goals and learned their associated policy (G2) the ultimate goal of this proposal is to develop an RL algorithm that takes as input a policy complexity parameter K—e.g. the maximum depth or number of nodes in a decision tree, the maximum number of prototypical situations to consider and so on—and returns the best policy in the given policy class. Advances compared to existing literature would be i) the ability to express the policy in terms of key relational attributes, instead of their absolute values, which will yield more compact and more general policies—addressing P1 and P2 respectively. ii) the ability to express policies in terms of intentions instead of low level actions which improves interpretability of the policy (P1) iii) the ability to scale to more complex visual inputs.

Impact. An algorithm returning interpretable policies can increase the adoption of RL by the industry. First, insights into an AI generated solution replacing a hand-coded one could be highly appreciated in an industrial context. Secondly, prior work has demonstrated that these simpler policies are amenable to stability analyses, at least when a model of the environment is known [3, 6], which can be critical in some industrial settings. Beyond the industry, this proposal can impact the research community: research on G1 could benefit the model-based RL community while research for G2 addresses long standing problems in the hierarchical RL community. The considered game domains in this proposal have 2D image inputs with fixed viewpoints. Extending our research to 3D scenes could reach robotics communities such as those focused on object manipulation.

3 Methodology and evaluation

The bulk of the research in this project will be conducted within the RL framework. While the policy structure is expected to have a simple form, internal models used by the algorithm can be arbitrarily complex and expressive. As such, we propose to cast the interpretable RL algorithm in the approximate policy iteration mold, with separate models for the policy and (a typically neural-based model) for the Q-function. For the representation learning part, the Ph.D. candidate is expected to at least have an affinity with computer vision and willingness to delve into its basic modules such as convolutional layers, attention modules and spatial transformers. For G1, the use of graph neural networks appears to be an appropriate choice. However, in our case the graph structure is unknown and uncovering it is in fact the main goal of G1. This is a challenging optimization problem—typically addressed using RL/random search methods such as REINFORCE—that is exacerbated in our case by the fact that the structure is dynamic. Overall, while the research goals give the impression that they can be solved in a sequential manner, in practice we expect the biggest challenge to be that of most RL approaches: the absence of i.i.d. data. Instead, model learning has to be mixed in with data collection making it harder to tackle our research goals in isolation on the Atari testbed. This challenge will be addressed by introducing an additional complementary task that focuses on only a subset of the research goals, as detailed below.

Evaluation. In addition to Atari games, we will consider the task of learning interpretable policies for the Rubik's cube puzzle. Rubik's cube was solved recently using search and deep RL methods [1] but it has interesting additional properties in the context of this proposal. First, there are known solutions for solving the puzzle from any initial configuration. These solutions are compact, human readable and can be easily memorized. They thus provide a perfect example of policy structures the AI should strive to autonomously learn and return. Secondly, a key aspect for solving the Rubik's cube is a set of action sequences (i.e. sub-policies) that can swap the position of some of the smaller cubes while leaving the remaining ones in place. This aligns well with the goals in G2 and the idea of finding 'disentangled sub-policies'. Finally, we can use the RL environment for the Rubik's cube developed by our team to access the full state information instead of learning from images. Overall, Rubik's cube is thus a good testbed to make progress specifically on G2, then on G3. On the other hand while some Atari games require complex chaining of skills, others like Pong are much simpler (e.g. a reasonable policy would be to simply keep the ball and the player's pad at the same height) and can be used to focus on G1 and G3.

In terms of evaluation metrics, progress on P1 (interpretability) will be ideally measured through user studies. If it is not possible to do so, or as a complementary metric, it is possible to use objective metrics when comparing interpretable policies of the same policy class. When comparing to RL-based methods, we should demonstrate that for a similar policy class and policy complexity we find policies with higher expected policy return. When comparing to imitation-based methods, we should demonstrate that we are able to find diverse strategies adapted for different policy complexities instead of mimicking the same strategy. To evaluate progress on P2 (generalization), we can use the setting studied in [4] to show that a policy expressed only w.r.t. the disentangled and relational features that matter for the task is robust to modifications of other irrelevant factors.

4 Position of the topic in Scool and in the scientific community at large

This topic is perfectly in-line and complementary to a set of research works on-going in Scool. It also tackles key questions studied by the scientific community.

In Scool, in the recent years, Ph. Preux and his students have been working on the coupling of discrete objects with continuous and statistical objects. The Ph.D. of Nathan Grinzstajn (2019–2022) is about the resolution of combinatorial optimization problems with uncertainty with RL. Considering that ambitious applications of RL require the use of knowledge about the task to solve, the internship (Spring-Summer 2021) and now the Ph.D. of Matheus Centa Medeiros (2021–2024) concerns the combination of symbolic knowledge in an RL agent. Combining discrete and continuous objects is also at the core of the post-doc of Mohit Mittal (2020–2022) as part of the HYAIAI Inria project, and that of Riccardo Della Vecchia (2021–2024) who studies how causality can provide an explanation of what an RL agent does in the Chist-Era CausalXRL project. Other projects are under submission (submission of a European project proposal in Fall 2021) or will soon exist with a company (Saint-Gobain) where these questions will also be investigated with post-docs. Applications currently studied in Scool are in the fields of patient follow-up (with CHU Lille) and agriculture in developing countries (with Cirad, CGIAR, and Bihar Agriculture University in India) will also benefit from this type of research.

The questions investigated in this Ph.D. are fully aligned with the topics of the HumAIn alliance.

More generally, the questions that will be studied during this Ph.D. have begun to be studied in the community but there is still a lot to do.

References

[1] Forest Agostinelli et al. "Solving the Rubik's cube with deep reinforcement learning and search". In: *Nature Machine Intelligence* (2019).

[2] R. Akrour, D. Tateo, and J. Peters. “Continuous Action Reinforcement Learning from a Mixture of Interpretable Experts”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[3] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. “Verifiable Reinforcement Learning via Policy Extraction”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[4] Ken Kansky et al. “Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics”. In: *International Conference on Machine Learning (ICML)*. 2017.

[5] Hyunjik Kim and Andriy Mnih. “Disentangling by Factorising”. In: *International Conference on Machine Learning (ICML)*. 2018.

[6] Mikel Landajuela et al. “Discovering symbolic policies with deep reinforcement learning”. In: *International Conference on Machine Learning (ICML)*. 2021.

[7] Guiliang Liu et al. “Learning Tree Interpretation from Object Representation for Deep Reinforcement Learning”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021.

[8] Francesco Locatello et al. “Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations”. In: *International Conference on Machine Learning (ICML)*. 2019.

[9] Tom Monnier et al. “Unsupervised Layered Image Decomposition into Object Prototypes”. In: *ICCV*. 2021.

[10] Abhinav Verma et al. “Programmatically Interpretable Reinforcement Learning”. In: *International Conference on Machine Learning (ICML)*. 2018.

[11] Angel Villar-Corrales and Sven Behnke. *Unsupervised Image Decomposition with Phase-Correlation Networks*. 2021.