Learning What to Attend to: Using Bisimulation Metrics to Explore and Improve Upon What a Deep Reinforcement Learning Agent Learns*

Nele Albers, Miguel Suau de Castro, and Frans A. Oliehoek

Delft University of Technology, Delft, Netherlands
{N.Albers, M.SuauDeCastro, F.A.Oliehoek}@tudelft.nl

Keywords: Deep Reinforcement Learning · Representation Learning · Bisimulation Metrics · Markovianity.

Recent years have seen a surge of algorithms and architectures for deep Reinforcement Learning (RL), many of which have shown remarkable success for various problems. Yet, little work has attempted to relate the performance of these algorithms and architectures to what the resulting deep RL agents actually learn, and whether this corresponds to what they should ideally learn. Such a comparison may allow for both an improved understanding of why certain algorithms or network architectures perform better than others and the development of methods that specifically address discrepancies between what is and what should be learned.

Ideal Representation. The concept of ideal representation we utilize is the Coarsest Markov State Representation (CMSR). We define this representation as one in which the Euclidean distances between states are proportional to how "behaviorally different" [2] those states are. Behavioral similarity thereby is measured by a specific bisimulation metric [1]. This bisimulation metric regards states as equivalent if and only if they have the same expected reward and transition distribution over all state equivalence classes for all actions. Moreover, if the parameters of two equivalent states are altered on a small scale, the metric distance between the states will stay small. Learning an internal state representation that is similar to the CMSR has several desirable theoretical properties:

- The CMSR is the smallest state representation that still allows for the prediction of the reward and next state [3].
- The CMSR does not distinguish states based on features that are irrelevant for predicting the next reward and internal state. Thus, a policy learned based on this representation generalizes to different values for such features.
- If a subset of the features required for predicting the reward and next internal state for a domain is sufficient for predicting the reward and next internal state after modifying the reward or the transition function, the CMSR for the original domain suffices to learn the Q-values of a thus modified domain.

* Full thesis available at http://resolver.tudelft.nl/uuid:2945dcc8-e7b9-4536-b9e7-074cfe86d3f9.

– Making the Euclidean distances between internal states proportional to their behavioral similarity renders the formed representation less sensitive to small estimation errors if the transition or reward functions are approximated.

Research Objective. It is hence *theoretically* desirable that deep RL agents learn the CMSR. Yet, we do not know to which extent deep RL agents learn the CMSR, and whether doing so is useful *in practice*. Thus, we look at the internal state representations learned by deep RL agents at various stages during training and under different training conditions, and compare them to the CMSR. Furthermore, to elucidate the practical usefulness of learning the CMSR, we contrast the learning speeds and consistencies and the generalization performances of neural networks with hidden-layer representations that differ in how similar to the CMSR they are, while controlling for other factors.

Contributions. We split our contributions into *methodological* and *experimental* ones. Our methodological contributions are as follows:

– We propose using correlation coefficients based on bisimulation metrics to measure how similar to the CMSR an internal state representation is. These correlation coefficients also allow to specifically determine whether an internal state representation is Markov with respect to the rewards or Markov with respect to the transitions[1].
– We introduce an auxiliary loss that pushes a neural network to learn an internal state representation that is similar to the CMSR in a network layer.

We further provide experimental contributions:

– We identify three overlapping learning phases that together make up the learning process of deep RL agents using model-free Q-learning agents as example. Thereby, it is during the second learning phase that internal state representations become increasingly similar to the CMSR. We also point out several factors that impact this learning process. The precise CMSR is not learned in any of our experiments.
– We show that learning a hidden-layer representation that is more similar to the CMSR *during* training can speed up the learning process and cause good solutions to be found more reliably.
– We demonstrate that learning a hidden-layer representation that is more similar to the CMSR *by the end of* training may lead to improved generalization to new irrelevant feature values. Creating such a representation also may enable better generalization to related domains with modified reward or transition functions, as long as the modifications do not render formerly irrelevant features relevant.

[1] A state representation that is *Markov with respect to the reward* is one in which knowledge of previous internal states does not lead to a more accurate prediction of the next reward [4]. The definition of *Markov with respect to the transition* proceeds analogously.

Acknowledgments. This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 758824 —INFLUENCE).

References

1. Ferns, N., Panangaden, P., Precup, D.: Metrics for finite markov decision processes. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. pp. 162–169. AUAI Press (2004)
2. Ferns, N., Precup, D.: Bisimulation metrics are optimal value functions. In: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence. pp. 210–219 (2014)
3. Givan, R., Dean, T., Greig, M.: Equivalence notions and model minimization in markov decision processes. Artificial Intelligence **147**(1-2), 163–223 (2003)
4. McCallum, R.: Reinforcement learning with selective perception and hidden state (1997)