About the Problems When Training Reinforcement Learning Agents for Verification Tasks

Timo P. Gros, Nicola J. Müller, and Verena Wolf

Saarland University, Saarland Informatics Campus, Saarbrücken, Germany
{timopgros,nmueller,wolf}@cs.uni-saarland.de

The intersection of (deep) reinforcement learning ((D)RL) and Verification is obvious: the foundational concept of deep reinforcement learning are Markov decision processes (MDPs) [8], a model also commonly used in the Verification community [2].

However, training RL agents for benchmarks commonly used in the Verification community comes with specific challenges that must be solved in order to use RL to reliably resolve nondeterminism.

Terminal-Only Reward. Training an RL agent to fulfill an arbitrary property yields an extremely sparse reward structure: positive when the goal is reached and zero elsewhere. Thus, a non-zero reward is only observed at the end of each episode, when a terminal state is reached. RL algorithms perform poorly when rewards are sparse [1,7], and this even harder setting institutes a challenge.

Undiscounted Objectives. Objectives originating from Verification may be unbounded or bounded with respect to time, but they typically do not measure how fast a goal condition was met [3,2]. For instance, consider the goal reachability probability. The objective is to maximize the probability of reaching the goal states, regardless of the number of steps taken to reach the goal.

In RL, it is common to use a discount factor $\gamma \in (0, 1)$ that is multiplied with each future reward, exponentially decreasing the weight of rewards the further they are in the future. Thus, short term rewards are more important and the RL agent is encouraged to reach the goal as soon as possible [8].

While using a discount factor of $\gamma = 1$ is permissible, it can introduce issues such as catastrophic forgetting or inhibit the agent from learning the task altogether

To utilize RL as a strategy to resolve nondeterminism in Verification benchmarks, these problems when using $\gamma = 1$ must be resolved.

Large Action Spaces. Common RL benchmarks can widely be controlled by a video game controller [5,6]. This clearly limits the action space.

However, the number of *actions*[1] occurring in the MDPs typically used in the Verification community is immense, significantly larger than the number of actions that can be represented by a video game controller [4].

Handling these large action spaces constitutes the third challenge.

In our work, we analyze the effects of the three stated challenges. We present how these issues are commonly handled and further present our attempts to tackle these challenges.

[1] From the Verification perspective, it might not even be obvious, what an action is: the *label* or the *transition*.

References

1. Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., Zaremba, W.: Hindsight experience replay. Advances in neural information processing systems **30** (2017)
2. Baier, C., Katoen, J.P.: Principles of model checking. MIT press (2008)
3. Grumberg, O., Clarke, E., Peled, D.: Model checking. In: International Conference on Foundations of Software Technology and Theoretical Computer Science; Springer: Berlin/Heidelberg, Germany (1999)
4. Hartmanns, A., Klauck, M., Parker, D., Quatmann, T., Ruijters, E.: The quantitative verification benchmark set. In: TACAS (1). pp. 344–350. LNCS 11427 (2019)
5. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing Atari with Deep Reinforcement Learning. arXiv preprint arXiv:1312.5602 (2013), accessed 15th September 2020
6. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M.A., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level Control through Deep Reinforcement Learning. Nature **518**, 529–533 (2015)
7. Pathak, D., Agrawal, P., Efros, A.A., Darrell, T.: Curiosity-driven explorationgrumberg1999model by self-supervised prediction. In: International conference on machine learning. pp. 2778–2787. PMLR (2017)
8. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. Adaptive computation and machine learning, The MIT Press, second edn. (2018)