Gregor Engels, Regina Hebig, Matthias Tichy (Hrsg.): SE 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023 53

Explainable Online Reinforcement Learning for Adaptive Systems

Felix Feit[1], Andreas Metzger[2], Klaus Pohl[3]

Abstract: This talk presents our work on explainable online reinforcement learning for self-adaptive systems published at the 3rd IEEE Intl. Conf. on Autonomic Computing and Self-Organizing Systems.

Keywords: Adaptation; Reinforcement Learning; Explainability; Interpretability

1 Presentation Summary

An adaptive system can automatically maintain its requirements in the presence of dynamic environment changes. Developing an adaptive system is difficult due to design time uncertainty, because how the environment will change at runtime and what precise effects adaptations will have on the running system are typically unknown at design time [We20].

Online reinforcement learning, i.e., employing reinforcement learning (RL) at runtime, is an emerging approach to realize self-adaptive systems in the presence of design time uncertainty. Online RL learns via actual operational data and thereby leverages feedback only available at runtime [Me22].

Deep RL algorithms represent the learned knowledge as a neural network. Compared with classical RL algorithms, Deep RL algorithms offer important benefits for adaptive systems. Deep RL can generalize over unseen inputs, it can handle continuous environment states and adaptation actions, and it can readily capture concept and data drifts [PMP20]. Yet, a fundamental problem of Deep RL is that learned knowledge is not represented explicitly. For a human, it is practically impossible to relate neural network parameters to concrete RL decisions. Figure 1 illustrates this problem by comparing how knowledge is represented.

Understanding RL decisions is key to (1) increase trust, and (2) facilitate debugging. Debugging is especially relevant for adaptive systems, because the reward function, which quantifies the feedback to the RL algorithm, must explicitly defined by developers, thus introducing a source for human error.

We introduce XRL-DINE to make Deep RL decisions for self-adaptive systems explainable [FMP22]. XRL-DINE enhances and combines explainable RL techniques from machine

[1] paluno, University of Duisburg-Essen, Essen, Germany, f.m.feit@gmail.com
[2] paluno, University of Duisburg-Essen, Essen, Germany, andreas.metzger@paluno.uni-due.de
[3] paluno, University of Duisburg-Essen, Essen, Germany, klaus.pohl@paluno.uni-due.de

CC BY SA

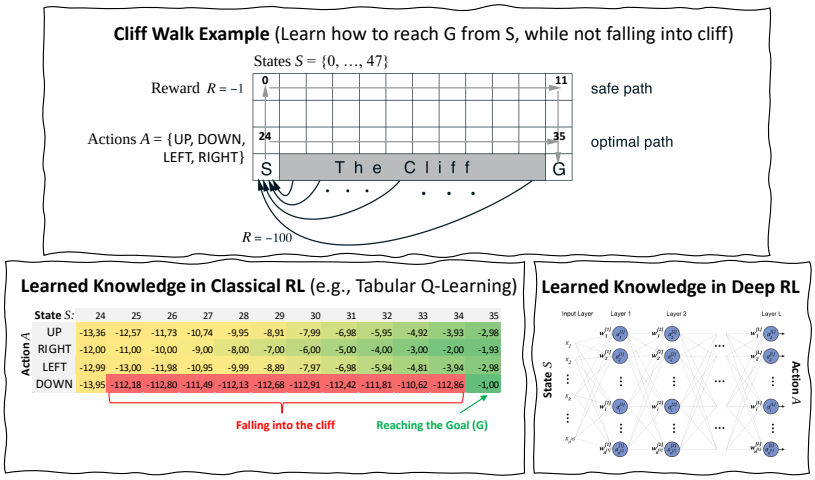

Abb. 1: Illustration of how learned knowledge is represented for Cliff Walk example from [SB18].

learning research. We present a proof-of-concept implementation of XRL-DINE, as well as qualitative and quantitative results that demonstrate the usefulness of XRL-DINE.

Data Availability. Source code and experimental data are available from `https://git.uni-due.de/rl4sas/xrl-dine`. The submission version of the original paper is available from `https://arxiv.org/abs/2210.05931`.

Literaturverzeichnis

[FMP22] Feit, Felix; Metzger, Andreas; Pohl, Klaus: Explaining Online Reinforcement Learning Decisions of Self-Adaptive Systems. In (Di Nitto, Elisabetta; Gerostathopoulos, Ilias; Bellman, Kirstie; Tomforde, Sven, Hrsg.): IEEE International Conference on Autonomic Computing and Self-Organizing Systems, ACSOS 2022, Virtual, September 19-23, 2022. IEEE, S. 51–60, 2022.

[Me22] Metzger, Andreas; Quinton, Clément; Mann, Zoltán Ádám; Baresi, Luciano; Pohl, Klaus: Realizing Self-Adaptive Systems via Online Reinforcement Learning and Feature-Model-guided Exploration. Computing, 2022.

[PMP20] Palm, Alexander; Metzger, Andreas; Pohl, Klaus: Online Reinforcement Learning for Self-adaptive Information Systems. In (Dustdar, Schahram; Yu, Eric; Salinesi, Camille; Rieu, Dominique; Pant, Vik, Hrsg.): 32nd International Conference on Advanced Information Systems Engineering, CAiSE 2020, Grenoble, France, June 8-12, 2020. Jgg. 12127 in LNCS. Springer, S. 169–184, 2020.

[SB18] Sutton, Richard S; Barto, Andrew G: Reinforcement learning: An introduction. MIT press, 2018.

[We20] Weyns, Danny: An Introduction to Self-adaptive Systems: A Contemporary Software Engineering Perspective. John Wiley & Sons, 2020.