Policy Presentation

Policy Summarization and Interactive Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Fabian Damken (fabian.damken@stud.tu-darmstadt.de)
March 2022, 10

Introduction

While intelligent agents outperform humans in various tasks, trust is hardly achievable as they do not exhibit their rationales. Also, training (deep) agents is extremely time-consuming. The focus of this report is twofold: exploring approaches in the field of *policy summarization* aiming to give a global overview of a policy and studying methods to incorporate human domain knowledge into training; *interactive learning*.

Policy Summarization

Explanations of single actions are not sufficient to grasp a policy completely if actions are executed at a high pace. Policy summarization aims to give a global overview. Two classes of methods for generating summaries are *trajectory extraction* and *textual summaries*.

Trajectory Extraction

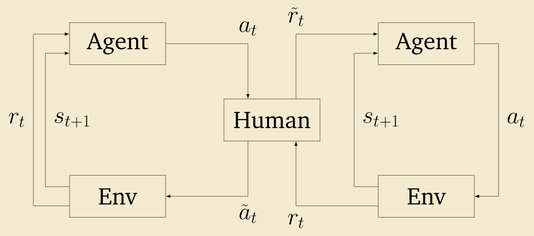
These methods work by selecting trajectories from a buffer that are *meaningful*. The trajectories are then presented to the human to provide explanations of the policy. The most challenging task is to decide which trajectories are *meaningful*.

Textual Summaries

These methods extract (natural) language explanation, mostly on single actions and not the global view. This family is not covered in detail.

Interactive Learning

If an agent exhibits bad behavior, there is no other option than to re-train. To remedy this, incorporating human feedback into learning can be helpful to actively influence the policy. Most methods in interactive learning use either *reward-* or *policy-shaping*.



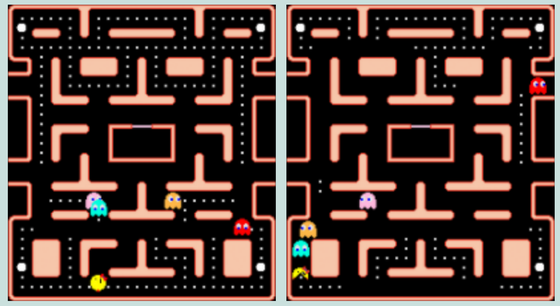
In reward-shaping, the reward is augmented by a human to reflect the goal and the agents learns using the augmented reward. In policy-shaping, the actions are judged directly before applying them to the environment. This report focuses on reward shaping as it is more common technique.

Conclusion

Both fields seem to evolve separately. However, it can be hard to grasp the policy an agent is actually executing and to give helpful feedback. One can imaging combined methods where the agent explains its rationale and the human can give feedback based on this explanation for future research.

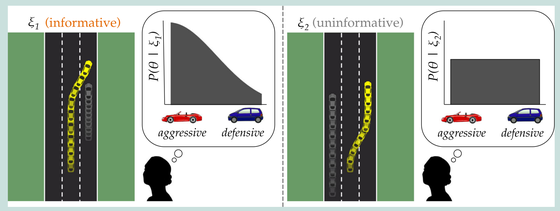
[1] Ofra Amir, Finale Doshi-Velez, and David Sarne. "Agent Strategy Summarization." In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 2018, pp. 1203–1207.
[2] Sandy H. Huang et al. "Enabling Robots to Communicate Their Objectives." In: *Autonomous Robots* 43.2 (Feb. 2019), pp. 309–326. ISSN: 1573-7527. DOI: 10.1007/s10514-018-9771-0.
[3] W. Bradley Knox and Peter Stone. "Interactively Shaping Agents via Human Reinforcement: The TAMER Framework." In: *Proceedings of the Fifth International Conference on Knowledge Capture*. K-CAP '09. New York, NY, USA: Association for Computing Machinery, Sept. 2009, pp. 9–16. ISBN: 978-1-60558-658-8. DOI: 10.1145/1597735.1597738.
[4] James MacGlashan et al. "Interactive Learning From Policy-Dependent Human Feedback." In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, July 2017, pp. 2285–2294.
[5] Matthew L. Olson et al. "Counterfactual State Explanations for Reinforcement Learning Agents via Generative Deep Learning." In: *Artificial Intelligence* 295 (June 2021), p. 103455. ISSN: 0004-3702. DOI: 10.1016/j.artint.2021.103455.

HIGHLIGHTS [1]



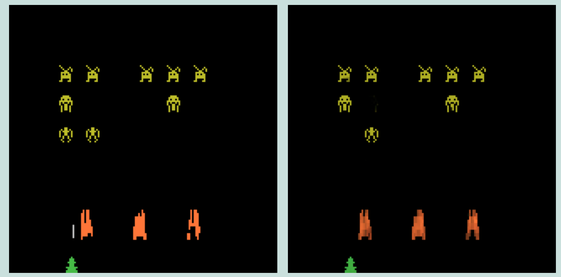
Trajectories are extracted from a buffer using the *state importance* $I^\pi(s) \coloneqq \max_{a \in \mathcal{A}} Q^\pi(s,a) - \min_{a \in \mathcal{A}} Q^\pi(s,a)$ as a measure of meaningfulness. Left is a state of low, right state is of high importance. To get a trajectory instead of a single state, a few states before and after are extracted, too. An extension, HIGHLIGHTS-DIV, additionally ensures variety in the extracted trajectories to tackle repetitive summaries.

Algorithmic Teaching [2]



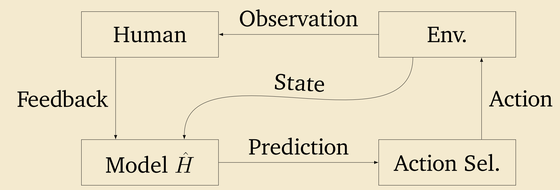
Instead of handcrafting an importance metric, it is assumed that humans perform Bayesian inference on the objective. States are extracted to maximize the posterior of the true objective. Left is an uninformative, right is an informative state. It was found that approximate inference outperforms exact inference, supporting the hypothesis that humans approximate.

Counterfactual Explanations [5]



Given two actions a and a' where the former was chosen by the policy, the minimal change in the state is shown that would have caused the agent to choose a', allowing to understand the agent's rationale. A summary is generated by pre-choosing action pairs and presenting the explanations.

TAMER [3]



A model $\hat{H}$ of the human reinforcement is learned. Previous states are weighted by a *credit* to reflect "recentness." Extensions were proposed to simultaneously learn from environmental reward.

COACH [4]

COACH accounts for the feedback to change when the policy changes. The feedback is interpreted as the *advantage* and the human serves as the "critic" in A2C.