A Policy-Graph Approach to Explain Reinforcement Learning Agents: A Novel Policy-Graph Approach with Natural Language and Counterfactual Abstractions for Explaining Reinforcement Learning Agents

Supplementary Material

A Experimental Environments

We here describe three environments tested and show their corresponding policy graphs generated by CAPS.

A.1 Cliffworld

Cliffworld is from example 6.5 in [24]. It is a standard gridworld with four actions in each state (up, down, right, left) which deterministically cause the corresponding state transitions (Figure 1). The reward is -1 for all transitions until the terminal state is reached but if the agent falls from the Cliff then the

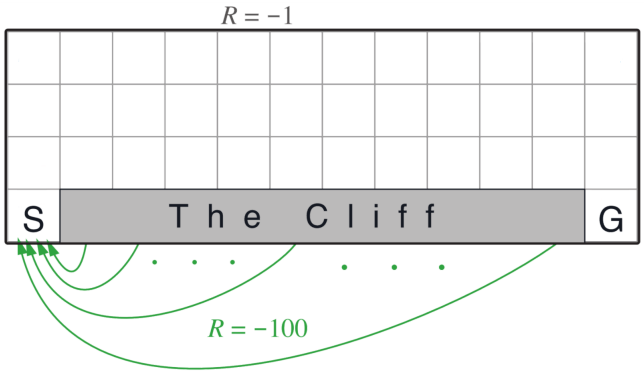

Figure 1: The state space of Cliffworld environment [24]

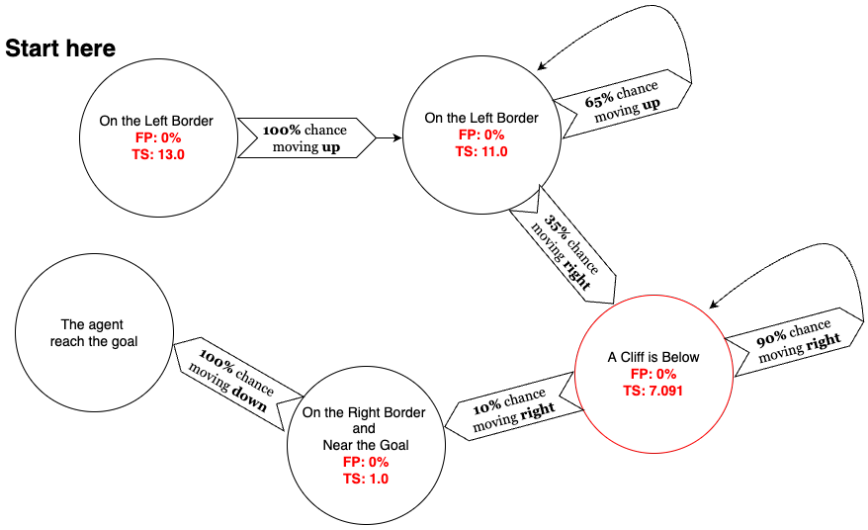

Figure 2: The CAPS Explanation for Cliffworld.

reward -100. The terminal state is in the bottom right corner, and the starting state is the bottom left corner [1].

The CAPS Explanation for Cliffworld can be interpreted as follows. The agent starts as the bottom left cell. It starts by moving up, and then moves right while the cliff is directly below it. It keeps moving right until it is close to the goal (the cell above the goal), then moves down.

A.2 GridWorld

We adapt Gridworld from [31], shown in Figure 3. In this environment, the agent starts in the bottom left corner, and the goal is to navigate to the top right corner using the actions up, left, down, and right. If the agent gets too close to the monster, or if it takes too many timesteps to reach the goal state, the agent fails and receives a large negative reward. The agent receives a reward of -1 for every timestep that is not associated with failure. In addition, when the agent is in the red area on the right of the grid, the monster will move towards the agent each timestep.

The CAPS Explanation for Gridworld (Figure 4) can be interpreted as follows. The agent starts at "at the start" and take steps to move toward the goal state. Since sometime the agent goes upward in "the normal path" or go rightward "leaving the start", the start state splits out into two paths with the same total number of timesteps. All failure probability stays with 0% due to no perturbation.

[1] The implementation of this environment in Open AI Gym is in `https://github.com/podondra/gym-gridworlds/blob/master/gym_gridworlds/envs/windy_gridworld_env.py`

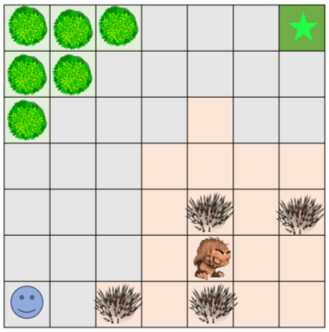

Figure 3: Gridworld environment [31]

A.3 Taxi

Taxi (Figure 6) is also a "gridworld" style environment where we have 4 designated locations (R, G, Y, B) and a passenger will be placed at one of these four locations, with their destination at one of this four locations, all randomly determined. The taxi will be placed randomly on the map, and the goal of the taxi is to pick up the passenger and drop them to the destination using the least timesteps possible. Due to the randomness of the environment in the starting position of the taxi and passenger, different trials would generate different graphs, thus, we fixed the locations of the taxi and passenger for the purpose of generating a consistent graph.

B User Study: More Results

C Experiments Hyperparameters

In 2, we show the hyperparameters we used in our expermints across the 5 environments. For the threshold t in Eq.8, we picked the top 15% of the states with least entropy and marked them as critical.

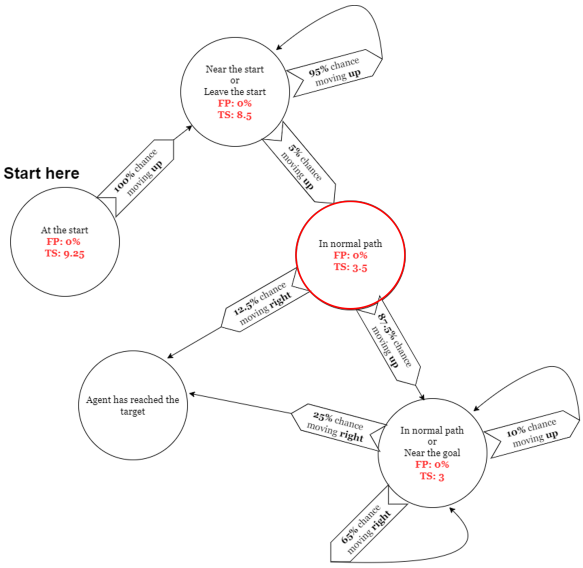

Figure 4: The CAPS explanation for Girdworld.

Figure 5: The Taxi environment

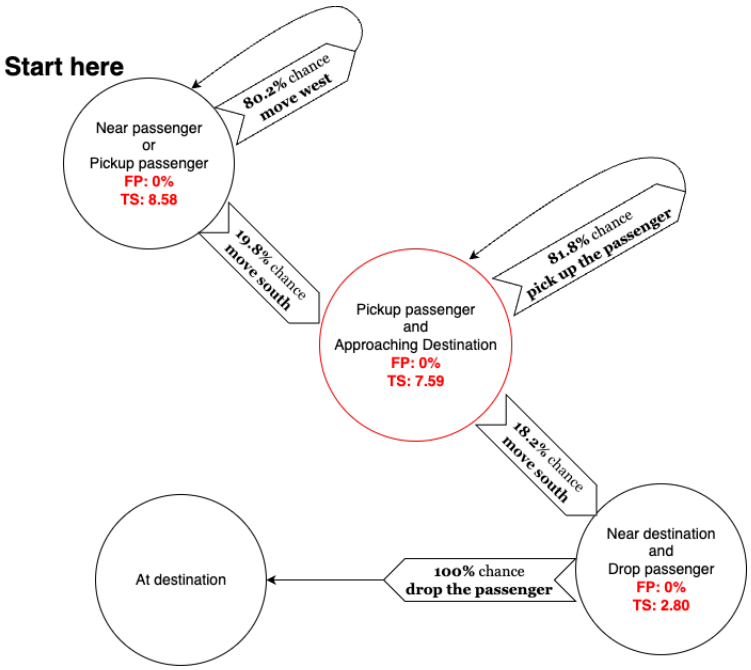

Figure 6: The CAPS explanation fro the Taxi environment

Method	Top 20 Slowest Responses				
	Time↓	S↑	A↑	S+A↑	S→A↑
Without graph	41s	-	0.05	-	-
Topin et.al [9]	159s	0.05	0.40	0.05	1.0 (1/1)
Zahavy et al. [8]	136s	0.45	0.45	0.10	0.22 (2/9)
CAPS (Optimal)	139s	0.85	0.95	0.80	0.94 (16/17)
Method	Top 30 Slowest Responses				
	Time↓	S↑	A↑	S+A↑	S→A↑
Without graph	38s	-	0.03	-	-
Topin et.al [9]	141s	0.03	0.33	0.03	1.0 (1/1)
Zahavy et al. [8]	125s	0.5	0.47	0.13	0.27 (4/15)
CAPS (Optimal)	126s	0.77	0.86	0.70	0.91 (21/23)

Table 1: Comparison of user-study results on (*Time*)–the total time spent on the task, accuracy of selecting the (S)–correct abstract state, (A)–correct action, (S+A)–correct abstract state and action, ($S \rightarrow A$)–correct action *after* correctly selecting abstract state.

Environment	CLTree Height	β in Eq.6
MountainCar	3	0.3
Blackjack	2	0.3
Cliffworld	3	0.3
Gridworld	3	0.3
Taxi	2	0.3

Table 2: The hyperparameters' values in our experiments

References

[1] Benbrahim, H., Franklin, J.A.: Biped dynamic walking using reinforcement learning. Robotics Auton. Syst. **22**, 283–302 (1997)

[2] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing Atari with Deep Reinforcement Learning (2013)

[3] Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J., Songhori, E., Wang, S., Lee, Y.-J., Johnson, E., Pathak, O., Bae, S., Nazi, A., Pak, J., Tong, A., Srinivasa, K., Hang, W., Tuncer, E., Babu, A., Le, Q.V., Laudon, J., Ho, R., Carpenter, R., Dean, J.: Chip Placement with Deep Reinforcement Learning (2020)

[4] Liu, N., Li, Z., Xu, J., Xu, Z., Lin, S., Qiu, Q., Tang, J., Wang, Y.: A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning, pp. 372–382 (2017). https://doi.org/10.1109/ICDCS.2017.123

[5] Peters, J., Schaal, S.: Policy gradient methods for robotics. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2219–2225 (2006). https://doi.org/10.1109/IROS.2006.282564

[6] Huang, S.H., Bhatia, K., Abbeel, P., Dragan, A.D.: Establishing appropriate trust via critical states. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2018)

[7] Hayes, B., Shah, J.: Improving robot controller transparency through autonomous policy explanation. 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI, 303–312 (2017)

[8] Zahavy, T., Ben-Zrihem, N., Mannor, S.: Graying the black box: Understanding dqns. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1899–1908. PMLR, New York, New York, USA (2016). https://proceedings.mlr.press/v48/zahavy16.html

[9] Topin, N., Veloso, M.: Generation of policy-level explanations for reinforcement learning. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, pp. 2514–2521. AAAI Press, ??? (2019). https://aaai.org/ojs/index.php/AAAI/article/view/4097

[10] Olson, M.L., Khanna, R., Neal, L., Li, F., Wong, W.-K.: Counterfactual state explanations for reinforcement learning agents via generative deep learning. Artificial Intelligence **295**, 103455 (2021). https://doi.org/10.1016/j.artint.2021.103455

[11] Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., Doshi-Velez, F.: Explainable reinforcement learning via reward decomposition. (2019)

[12] Aniek Markus, J.K., Rijnbeek, P.: The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics **113**, 103655 (2021). https://doi.org/10.1016/j.jbi.2020.103655

[13] Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: Explainable reinforcement learning through a causal lens. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, pp. 2493–2500. AAAI Press, ??? (2020). https://aaai.org/ojs/index.php/AAAI/article/view/5631

[14] Liu, B., Xia, Y., Yu, P.S.: Clustering via decision tree construction. (2004)

[15] Schulman, J., Filip Wolski, A.R. Prafulla Dhariwal, Klimov, O.: Proximal Policy Optimization Algorithms (2017)

[16] Volodymyr Mnih, D.S. Koray Kavukcuoglu, Alex Graves, I.A., Wierstra, D., Riedmiller, M.: Playing Atari with Deep Reinforcement Learning (2013)

[17] Liu, G., Schulte, O., Zhu, W., Li, Q.: Toward Interpretable Deep Reinforcement Learning with Linear Model U-Trees (2018)

[18] van der Waa, J., van Diggelen, J., van den Bosch, K., Neerincx, M.: Contrastive Explanations for Reinforcement Learning in terms of Expected Consequences (2018)

[19] Iyer, R.R., Li, Y., Li, H., Lewis, M., Sundar, R., Sycara, K.P.: Transparency and explanation in deep reinforcement learning neural networks. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018)

[20] Yang, Z., Bai, S., Zhang, L., Torr, P.H.S.: Learn to Interpret Atari Agents (2019)

[21] Greydanus, S., Koul, A., Dodge, J., Fern, A.: Visualizing and understanding Atari agents. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1792–1801. PMLR, ??? (2018)

[22] Amir, D., Amir, O.: Highlights: Summarizing agent behavior to people. In: AAMAS (2018)

[23] McCalmon, J., Le, T., Alqahtani, S., Lee, D.: Caps: Comprehensible abstract policy summaries for explaining reinforcement learning agents. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. AAMAS '22, pp. 889–897. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2022)

[24] S.Sutton, R., Precup, D., Singh, S.: Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence **112**, 181–211 (1999)

[25] Moore, A.: Variable resolution reinforcement learning. (1995)

[26] Rodgers, J., Nicewander, A.: Thirteen ways to look at the correlation coefficient. American Statistician - AMER STATIST **42**, 59–66 (1988). https://doi.org/10.1080/00031305.1988.10475524

[27] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)

[28] Yash Goyal, J.E.D.B.D.P. Ziyan Wu, Lee, S.: Counterfactual visual explanations. In: ICML, pp. 2376–2384 (2019). http://proceedings.mlr.press/v97/goyal19a.html

[29] Uesato Jonathan, S.C.E.T. Kumar Ananya, Ruderman, K.K. Avraham Anderson, Dvijotham ad Heess, N.K., Pushmeet: Rigorous Agent Evaluation: An Adversarial Approach to Uncover Catastrophic Failures. arXiv (2018). https://doi.org/10.48550/ARXIV.1812.01647. https://arxiv.org/abs/1812.01647

[30] Abolfathi, E.A., Luo, J., Yadmellat, P., Rezaee, K.: Coachnet: An adversarial sampling approach for reinforcement learning. arXiv preprint arXiv:2101.02649 (2021)

[31] van der Waa Jasper, B.K.v.d. van Diggelen Jurriaan, Mark, N.: Contrastive explanations for reinforcement learning in terms of expected consequences (2018). https://doi.org/10.48550/ARXIV.1807.08706

[32] Brockman, G., Cheung, V., Ludwig Pettersson, J.S.J.T. Jonas Schneider, Zaremba, W.: OpenAI Gym (2016)