Toward Robust Policy Summarization

Extended Abstract

Isaac Lage
Harvard University

Daphna Lifschitz
Technion - Israel Institute of Technology

Finale Doshi-Velez
Harvard University

Ofra Amir
Technion - Israel Institute of Technology

ABSTRACT

AI agents are being developed to help people with high stakes decision-making processes from driving cars to prescribing drugs. It is therefore becoming increasingly important to develop "explainable AI" methods that help people understand the behavior of such agents. Summaries of agent policies can help human users anticipate agent behavior and facilitate more effective collaboration. Prior work has framed agent summarization as a machine teaching problem where examples of agent behavior are chosen to maximize reconstruction quality under the assumption that people do inverse reinforcement learning to infer an agent's policy from demonstrations. We compare summaries generated under this assumption to summaries generated under the assumption that people use imitation learning. We show through simulations that in some domains, there exist summaries that produce high-quality reconstructions under different models, but in other domains, only matching the summary extraction model to the reconstruction model produces high-quality reconstructions. These results highlight the importance of assuming correct computational models for how humans extrapolate from a summary, suggesting human-in-the-loop approaches to summary extraction.

KEYWORDS

Explainable AI; Policy Summarization

ACM Reference Format:
Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. 2019. Toward Robust Policy Summarization. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

AI agents are being developed to help people with high stakes decision-making processes from driving cars to prescribing drugs. These agents offer the promise of learning to act optimally from data, but their behavior can be opaque to the human users who interact with them. Involving a human user in the evaluation process can help guard against potential harms due to failings in agent training, and elucidate the strengths of successful agents.

To support people's understanding of agent policies, methods for extracting summaries of an agent's policy, i.e. informative collections of agent decisions consisting of state-action pairs, have been proposed [3]. One approach to summarization relies on heuristics for diversity or state importance [2, 7] . Another approach assumes a computational model of how humans will generalize from a summary, and uses this model to optimize summaries to aid in reconstructing the policy [8]. Specifically, Huang *et al.* [8] assumed people would employ reasoning akin to inverse reinforcement learning (IRL) to understand an agent's objective, and extract a summary that allows for accurate approximation of the agent's reward. But people may also do imitation learning (IL), mapping from states directly to actions. The cognitive science literature shows that users apply different computational models in different situations while planning [6], suggesting that they may do the same when inferring an agent's policy from a summary.

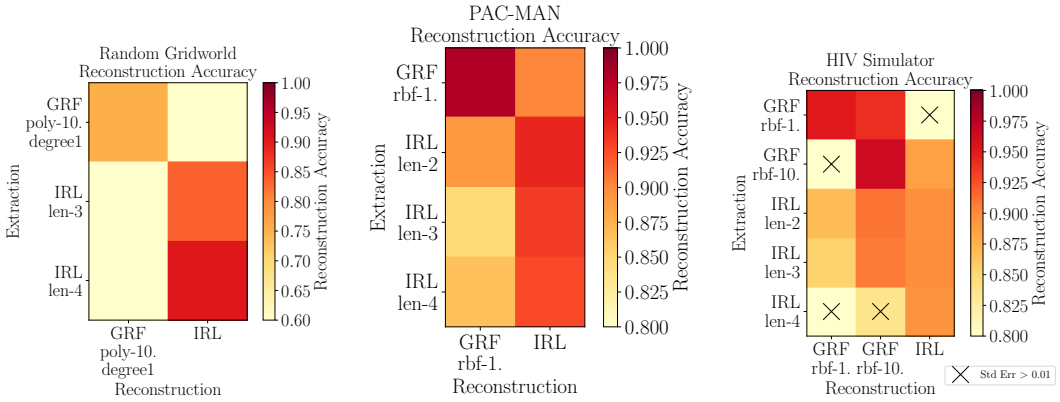
We investigate the question of whether summary quality is robust to a mis-match between the model assumed during summarization and the user's true model. We run computational simulations exploring whether summaries extracted assuming that people do IRL allow an IL model to reconstruct the policy, and vice versa. Our results show that in some domains, assuming the correct reconstruction model during summary extraction is necessary to accurately reconstruct an agent's policy using that summary, but that in other domains, we can extract summaries that are robust to misspecification of the reconstruction model during summary extraction. This highlights the importance of considering how humans will extrapolate from summaries when selecting which agent behaviors (state-action pairs) to show.

2 METHODOLOGY

We assume summaries in the form a set of state-action pairs $T = \langle\langle s_1, a_1\rangle, ..., \langle s_k, a_k\rangle\rangle$ from a batch of agent demonstrations, chosen to maximize the quality of a simulated user's understanding of the agent's policy derived from these examples. The models we use to simulate how users infer agents' policies from summaries are variants of IRL and IL, described below with the corresponding procedures we employ for extracting summaries.

IRL Given a collection of trajectories, Inverse Reinforcement Learning (IRL) extracts a reward function such that the optimal policy with respect to those rewards matches the demonstrated behavior [9]. We use the Maximum Entropy IRL (Max-Ent) model [11] as a proxy of how people may extract such reward functions given a collection of trajectories. To generate the summaries we use the algorithmic teaching approach presented in [4] to extract a set of trajectories that optimally reconstruct the true policy. We modified the algorithm to extract a fixed budget by terminating after k states, or by randomly adding trajectories when the algorithm terminates

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.



(a) Random Grid World **(b) Pacman** **(c) HIV Simulator**

Figure 1: Accuracy averaged over 20 restarts for every combination of reconstruction model (columns) and summary extraction model (rows). The number for IL models corresponds to length scale, and for IRL models to trajectory lengths. All reconstruction models reconstruct the policy most accurately with the matched summary extraction model. For the gridworld, no summary extraction models are robust to mis-match with the reconstruction model; for PAC-MAN and HIV, some summary extraction models are more robust to mis-match with the reconstruction model than others.

with less than k states. IRL-based extraction of summaries was also explored in [8].

IL An alternative model for human extrapolation, rather than identifying rewards and planning against them, tries to directly mimic the agent. Imitation learning captures the notion that people may infer agent policies based on the actions an agent has taken in similar states, with no concept of reward or goal. In our experiments, we use the Gaussian random field (GRF) model and the active learning algorithm described in Zhu *et al.* [10] to find the set of size k that maximizes accuracy on states not included in the summary. To our knowledge, IL-based summary extraction has not been previously studied.

Domains We use the random grid world defined in [4]; a 6x7 PAC-MAN grid with a single food pellet in the middle, a wall surrounding it on 3 sides, and a ghost that moves towards PAC-MAN deterministically [1]; and the HIV simulator described in [1]. We derive the policies for the domains respectively with: value iteration using discount factor 0.95; so the agent moves in the direction of the nearest food that does not result in a collision with the ghost; and using fitted Q iteration as in [5] with a 0.05 initial state perturbation. For PAC-MAN, we derived distinct feature sets for IL and IRL based on distance to food, walls and ghost attacks. For HIV, we discretized states using K-Means clustering with 100 clusters.

Metric We use accuracy for all unique states not in the summary.

3 RESULTS

In Figure 1, we present results for hyperparameters (listed in figure) and summary sizes (gridworld: 48; PAC-MAN: 24; HIV: 24) that outperform other hyperparameter settings on reconstruction quality with summaries optimized for them.

Across all datasets, all models reconstruct the policy most accurately when the summary is extracted using the same model. In Figure 1, different reconstruction models perform better in different domains (IL is more accurate for HIV and PAC-MAN; IRL is more accurate for gridworld), but a fixed reconstruction model always produces the highest quality reconstruction with the summary optimized for it. This suggest that both the IL and IRL methods extract summaries that allow the model assumed during summary extraction to reconstruct the policy accurately.

In some domains, using any model of human computation other than the correct one during summary extraction leads to poor reconstruction. In the gridworld domain, the IL model cannot reconstruct the policy well with either IRL summary, and the IRL model cannot reconstruct the policy well with the IL summary. This suggests that sometimes, knowing the reconstruction model is necessary to extract summaries that produce high quality reconstructions.

In other domains, some summaries are more robust to a mismatch between summarization and reconstruction models than others. In PAC-MAN, the summary extracted with IRL and trajectory length 2 allows the GRF model to reconstruct the policy more accurately than the summaries generated with other choices of trajectory length. In the HIV simulator, the summary extracted with the GRF model with length scale 10 allows the IRL model to reconstruct the policy more accurately than the GRF model with length scale 1. The summaries extracted with the IRL model with trajectory lengths 2 and 3 allow both GRF models to reconstruct the policy more accurately than the summary extracted with the IRL model and trajectory length 4. This suggests that there are domains where some summaries are more robust to a mis-match between summarization and reconstruction model than others.

4 ACKNOWLEDGMENTS

The authors acknowledge a Google Faculty Research Award and the JP Morgan AI Faculty Research Award. IL is supported by NIH 5T32LM012411-02.

[1] http://ai.berkeley.edu/project_overview.html

REFERENCES

[1] BM Adams, HT Banks, M Davidian, Hee-Dae Kwon, HT Tran, SN Wynne, and ES Rosenberg. Hiv dynamics: modeling, data analysis, and optimal treatment protocols. *Journal of Computational and Applied Mathematics*, 184(1):10–49, 2005.

[2] Dan Amir and Ofra Amir. Highlights: Summarizing agent behavior to people. In *Proc. of the 17th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2018.

[3] Ofra Amir, Finale Doshi-Velez, and David Sarne. Agent strategy summarization. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1203–1207. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[4] Daniel S. Brown and Scott Niekum. Machine teaching for inverse reinforcement learning: Algorithms and applications. *CoRR*, abs/1805.07687, 2018.

[5] Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Decision and Control, 2006 45th IEEE Conference on*, pages 667–672. IEEE, 2006.

[6] Samuel J Gershman. Reinforcement learning and causal models. *The Oxford handbook of causal reasoning*, page 295, 2017.

[7] Sandy H. Huang, Kush Bhatia, Pieter Abbeel, and Anca D. Dragan. Leveraging critical states to develop trust. In *RSS 2017 Workshop: Morality and Social Trust in Autonomous Robots*, 2017.

[8] Sandy H. Huang, David Held, Pieter Abbeel, and Anca D. Dragan. Enabling robots to communicate their objectives. *Autonomous Robots*, 43(2):309–326, Feb 2019.

[9] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[10] Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. *ICML 2003 Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65, 2003.

[11] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.