Goal-Driven Active Learning

JAAMAS Track

Nicolas Bougie
The Graduate University for Advanced Studies (Sokendai),
National Institute of Informatics
Tokyo, Japan
nicolas-bougie@nii.ac.jp

Ryutaro Ichise
National Institute of Informatics,
The Graduate University for Advanced Studies (Sokendai)
Tokyo, Japan
ichise@nii.ac.jp

ABSTRACT

Despite recent breakthroughs for learning a rich set of behaviors in simulated tasks, reinforcement learning agents are not yet in widespread use in the real world where rewards are naturally sparse. In fact, efficient exploration remains a key challenge in sparse-reward tasks as it requires quickly finding informative and task-relevant experiences. While cloning behaviors provided by an expert is a promising approach to the exploration problem, learning from a fixed set of demonstrations may be impracticable due to lack of state coverage or distribution mismatch - when the learner's goal deviates from the demonstrated behaviors. Moreover, we aim to obtain a policy that can accomplish a variety of goals guided by the same set of demonstrations (i.e. without additional human effort). We present a goal-conditioned method that leverages very small sets of goal-driven demonstrations to significantly accelerate learning. Crucially, we present the concept of active goal-driven demonstrations to query the demonstrator only in hard-to-learn and uncertain regions of the state space. We evaluate our framework on a set of robot control tasks. Our method outperforms prior imitation learning approaches in most of the tasks in terms of data efficiency and scores while reducing the amount of human effort.

KEYWORDS

Deep Reinforcement Learning; Imitation Learning; Goal-Conditioned Learning; Active Learning

ACM Reference Format:
Nicolas Bougie and Ryutaro Ichise. 2022. Goal-Driven Active Learning: JAAMAS Track. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

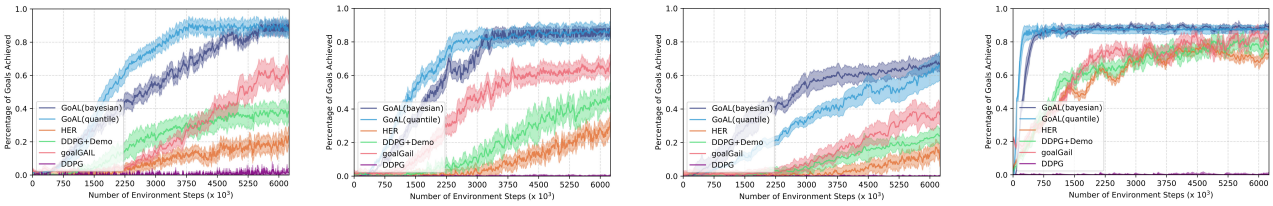
Reinforcement learning (RL) has shown impressive advances in a plethora of simulated tasks, including game-playing [8] or robot control [7]. On the other hand, many real-world problems involve rewards that are sparse or delayed, which limits the applicability of RL. As a result, such approaches require a large number of interactions to reach decent performance, which is often intractable. Therefore, achieving efficient exploration is a key challenge to expand the possible applications of RL.

Multiple approaches were proposed to achieve better explorative policies. One strategy is goal-conditioned learning, a form of self-supervision that constructs a goal-conditioned policy [6, 11]. The objective is to reach any goal upon demand. This idea was extended in Hindsight Experience Replay (HER) [2] to artificially generate additional transitions by relabeling goals seen along the state trajectory. However, these algorithms might still produce ineffective learning of complex policies - it may require a large amount of data to capture complex or far-away goals.

Since it is often unrealistic to expect an end-to-end reinforcement learning system to rapidly succeed with no prior assumptions about the domain, multiple methods have introduced prior knowledge into reinforcement learning systems. The most common form of external supervision is imitation learning. Imitation learning seeks to learn tasks from demonstrated state-action trajectories [1, 10]. For instance, Deep Q-learning from Demonstrations [5] improves initial performance by pre-training the policy with demonstrations. However, learning from human demonstrations suffers from three problems. (1) It is hard to obtain a broad state coverage of task-relevant regions from trajectories demonstrated without specific goals. (2) It usually has an abundance of irrelevant or redundant information. In this sense, imitation learning puts more burden on humans than just providing insights about hard-to-learn regions of the environment. (3) It assumes that the learner's goal matches the expert's demonstrated behaviors. Besides, most imitation learning algorithms learn policies that achieve a single task.

In this work [3], we propose an active goal-conditioned approach that drastically reduces expert workload by incrementally requesting partial demonstrations towards specific goals, *goal-driven demonstrations*. In contrast with standard demonstrations, goal-driven demonstrations do not aim to demonstrate the overall task or all possible situations. Instead, goal-driven demonstrations fulfill particular goals that are actively chosen based on the agent's knowledge about the task. To do so, the present approach allows an agent to jointly identify states where feedback is most needed and communicate for specific domain knowledge throughout the training. Namely, goal-driven demonstrations are actively queried based on the agent's confidence and the ability of the agent to reach the goal being pursued. We build and compare two techniques to estimate the agent's confidence: 1) Bayesian-confidence, 2) quantile-confidence; and study a relabeling strategy that extracts additional information from the demonstrated trajectories. In addition, we present a form of Q-filter that allows the agent to outperform the demonstrator. Overall, this scheme constitutes a novel perspective for robotic task learning and could help to expand the possible applications of RL.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.



(a) Hand Manipulate Block (b) Hand Manipulate Egg (c) Hand Manipulate Pen (d) Hand Reach

Figure 1: Performance (mean±std) on ShadowHand tasks averaged over 10 runs for our method using different estimations of the agent's confidence (GoAL(bayesian), GoAL(quantile)) and baselines DDPG [7], HER [2], DDPG+Demo [9], and goalGail [4].

2 PROPOSED FRAMEWORK

The proposed framework provides us a mechanism to mitigate the above-mentioned problems by incrementally querying goal-driven demonstrations. Namely, our approach introduces human feedback into goal-conditioned learning via HER. The agent receives feedback in the form of short goal-driven demonstrations—the tutor is requested to reach a specific goal. We decide how to query goal-driven demonstrations based on the agent's needs and the expected value of information of the query, drastically reducing the number of required demonstrations.

The framework consists of five steps. We first collect a trajectory based on the goal being pursued by running a goal-conditioned agent. Second, the agent decides whether it should query a goal-driven demonstration to the demonstrator by evaluating the agent's confidence. In this work we present and compare two techniques to estimate the agent's confidence: 1) Bayesian-confidence, 2) quantile-confidence. Given such a query, the teacher is requested to reach a specific goal that has been chosen by the agent. Third, after each query, we perform expert relabeling to artificially generate more expert data. Expert relabeling is a type of data augmentation on the provided goal-driven demonstrations. Fourth, an imitator policy is then trained to imitate the demonstration data. Note that in order to allow the agent to significantly outperform the demonstrator, we use a Q-filter function [9] in a goal-conditioned setting, which we extend to take into account the gap between "optimal" and "sub-optimal" transitions. Finally, we augment the policy loss with an extra objective that aims to mimic the demonstrated behaviors. Note that the transitions used to train the policy are generated following a similar strategy as in HER [2], except that we modified the goal sampling to take advantage of the demonstrations. This process continues until the task is mastered.

3 EXPERIMENTS AND RESULTS

Via a wide range of simulations, the paper analyses and compares the proposed framework with several baselines on eight robotic tasks implemented in MujoCo [12]. Simulations allow us to systematically evaluate the performance of our system under different hypotheses about the teaching conditions (e.g. low query budget), and to test its limits. For instance, we evaluate the robustness of our framework against noisy guidance data and erroneous teaching signals. The experimental results of the simulations can be summarized as follows:

Ideal Case When teaching signals are correct, our method improves the convergence speed, and in some tasks the final performance with respect to baselines approaches, as shown in Figure 1. Please note that we compare the average learning performance of our method that uses two different mechanisms to evaluate the agent's confidence when deciding to make a query.

Better-than-expert Performance The experimental results show that the proposed Q-filter technique allows the agent to outperform the teacher by discarding sub-optimal expert transitions, yielding better-than-expert performance.

Erroneous instructions We study how our agent performs when imperfect guidance is generated by the demonstrators by adding normal noise. The proposed method is reasonably robust to noise in the demonstrations, and hence a non-expert can provide a feedback signal to the agent.

Generalization to unseen goals When training our agent on a set of goals (i.e., not contained within the provided guidance), the agent can generalize the provided guidance to unseen goals with a slight loss in the performance.

Low Query Budget Our method leverages a small amount of demonstrations that cover task-relevant regions of the state space, which entails that it remains effective with a low query budget. In addition, the present method outperforms the baselines by a large margin under the same query budget.

4 CONCLUSION

This work presents a method for actively teaching an agent with goal-driven demonstrations to both learn more effectively and efficiently. Goal-driven demonstrations do not intend to demonstrate the overall task, but help the agent to fulfill particular intermediate goals when it struggles. In contrast with traditional imitation learning approaches where the agent passively accesses to the demonstration data, our approach actively decides when to request goal-driven demonstrations based on the confidence of the agent. In addition, we show how to generate additional expert data by relabeling goal-driven demonstrations. As a result, this novel form of human guidance is less expensive and more intuitive than pure demonstrations, while ensuring that the provided knowledge match the agent's needs, hence escaping the known "distribution mismatch" issues of prior work. A promising research direction is to replace the human trainer with another source of guidance. For instance, if demonstrations are not available, one solution is to reuse successful rollouts as demonstration data.

REFERENCES

[1] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the international conference on Machine learning*. 1–8.

[2] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. In *Advances in neural information processing systems*. 5048–5058.

[3] Nicolas Bougie and Ryutaro Ichise. 2021. Goal-driven active learning. *Autonomous Agents and Multi-Agent Systems* 35, 2 (2021), 1–29.

[4] Yiming Ding, Carlos Florensa, Pieter Abbeel, and Mariano Phielipp. 2019. Goal-conditioned imitation learning. In *Advances in Neural Information Processing Systems*. 15298–15309.

[5] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, Gabriel Dulac-Arnold, John Agapiou, Joel Z Leibo, and Audrunas Gruslys. 2018. Deep Q-learning from Demonstrations. In *Proceedings of the Annual Meeting of the Association for the Advancement of Artificial Intelligence*.

[6] Leslie Pack Kaelbling. 1993. Learning to achieve goals. In *Proceedings of the International Joint Conferences on Artificial Intelligence*. 1094–1098.

[7] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint:1509.02971* (2015).

[8] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, and Georg. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[9] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. 2018. Overcoming exploration in reinforcement learning with demonstrations. In *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, 6292–6299.

[10] Stefan Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences* 3, 6 (1999), 233–242.

[11] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. 2015. Universal value function approximators. In *Proceedings of the International conference on machine learning*. 1312–1320.

[12] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *Proceedings of the International Conference on Intelligent Robots and Systems*. 5026–5033.