Do As You Teach: A Multi-Teacher Approach to Self-Play in Deep Reinforcement Learning

Extended Abstract

Chaitanya Kharyal
Microsoft
Hyderabad, India
chaitanyajee@gmail.com

Tanmay Sinha
Microsoft Research
Bangalore, India
tanmaysinha18@gmail.com

Sai Krishna Gottipati
AI Redefined
Montreal, Canada
sai@ai-r.com

Fatemeh Abdollahi
University of Alberta
Alberta, Canada
fabdolla@ualberta.ca

Srijita Das
University of Alberta
Alberta, Canada
srijita1@ualberta.ca

Matthew E. Taylor
University of Alberta
AI Redefined
Alberta, Canada
matthew.e.taylor@ualberta.ca

KEYWORDS

Self-Play; Curriculum Learning; Reinforcement Learning

ACM Reference Format:
Chaitanya Kharyal, Tanmay Sinha, Sai Krishna Gottipati, Fatemeh Abdollahi, Srijita Das, and Matthew E. Taylor. 2023. Do As You Teach: A Multi-Teacher Approach to Self-Play in Deep Reinforcement Learning: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

The future of industrial automation is hinged on the ability of the industrial robots to precisely finish the tasks designated for them [5]. These tasks are usually specified in terms of a state the robot is required to reach (i.e., a goal state). Goal-conditioned reinforcement learning [7, 8] is an emerging sub-field that trains policies with goal inputs. This enables the agent to generalize to new unseen goals, learn multiple complex tasks and acquire new skills along the way.

Training a goal conditioned reinforcement learning (RL) agent in sparse-reward environments that could generalize well to other unseen goals has been a long lasting challenge (owing to factors like catastrophic forgetting and poor credit assignment). While several exploration based methods are proposed [1, 2, 4], they all try to optimize for a specific objective (e.g, information theoretic based reward, intrinsic reward, count-based reward, etc.). It remains unclear whether these align with the actual objective of the goal-conditioned agent and if the same algorithm can work on a wide range of environments.

In this work, we propose a novel algorithm to generate an automatic curriculum of increasingly challenging goals set by one or more teachers that act in the environment for a goal-conditioned student agent. The intuition behind our proposed algorithm is that multiple teachers would cover larger parts of the state-space, thus leading to better approximation of the goal distribution and suggesting diverse goals.

2 MULTI-TEACHER CURRICULUM LEARNING

We train two types of agents — *teacher(s)* and a *student*. Similar to Asymmetric Self Play (ASP) [6, 9], in each teacher-student interaction, we sample a starting state $s_0 \in \mathcal{S}$. From this state, the teacher's policy $\pi_T(a \mid s)$ selects actions over time and eventually reaches its final state, $g \in \mathcal{S}$. Starting from the same state s_0, the student, with its goal conditioned policy $\pi_S(a \mid s, g)$, interacts with the environment and tries to reach the goal state set by the teacher, g_t. This entire course of interaction is referred to as a teacher-student rollout

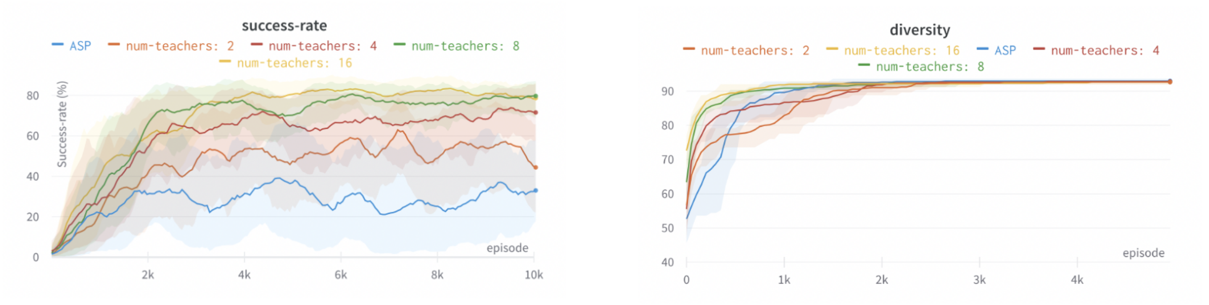
2.1 Problem setup

Given: A single student agent S and a set of teacher agents $\{T_1, T_2, \cdots, T_N\}$
Objective: Train the student agent S so that it learns an optimal goal-conditioned policy that generalizes to unseen goals
Assumptions: Our method does not assume the teachers have any expertise and must learn to propose goals on the fly, influenced by the student's learning performance or acquired skill. All the agents in our proposed approach are learning agents starting from scratch (without any pre-training).

2.2 Algorithm and methodology

The multi-teacher ASP algorithm (MT-ASP) is detailed in Algorithm 1. We denote the n teacher agents as T_1, T_2,... T_N and the student agent as S. Consequently we represent the parameters of actor and critic networks of teacher agents with $\theta_{T_1}, \cdots, \theta_{T_N}$ and that of the student agent with θ_S. An episode consists of $(N \cdot m)$ student-teacher rollouts. In every episode, we do a rollout of a single teacher agent m times wherein the teacher agent interacts with the environment to set a goal (line 5) followed by a rollout of the student agent, where the student attempts to reach the goal set by the previous teacher (line 7). After $(N \times m)$ student-teacher rollouts, we update the parameters of each teacher agent θ_{T_i} based on the actor and critic loss functions and parameters of the student agent θ_S is updated with both a behavior cloning loss and the standard RL loss as described later. (lines 10-11). Every teacher and student agent uses its own rollout data (experience) to update their respective parameters.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.



(a) Success rate on random goals vs episode

(b) Cumulative % of states visited by teacher vs episode

Figure 1: Success rate and cumulative % of states visited for Fetch Reach domain

Algorithm 1: Multi Teacher Asymmetric Self-Play

```
Data: $N, m$ ; //Number of teacher agents, multiplier
Data: $\theta_{T_1}, \cdots, \theta_{T_N}, \theta_S$ ;  //Parameters for the agents
for episode = $1, 2, \cdots$ do
    for trial$(i) = 1, 2, \cdots, N \cdot m$ ;              //Rollouts
     do
        $k = \lfloor i/m \rfloor$;
        $k^{th}$ teacher sets goal;
        if goal is valid then
            Student tries to achieve the goal;
    for $i = 1, 2, \cdots, N \cdot m$ do
        $k = \lfloor i/m \rfloor$;
        Update $\theta_{T_k}$ ;    //TD3 Loss using $T_k$'s replay
        Update $\theta_S$ ;                 //TD3 and BC Loss
```

Reward: We assign rewards to both the teacher and student agents based on whether the student is able to reach the goal set by a teacher. If the student reaches the goal set by a teacher, the particular teacher gets a fixed negative reward and the student gets a fixed positive reward. Otherwise, when the student does not reach the goal, the teacher gets a fixed positive reward and the student gets a reward of 0. Furthermore, the notion of invalid (unwanted) goals can be added to this reward structure by giving the teacher a large negative reward for setting an invalid goal. We have tested the invalid goal hypothesis on the fetch reach environment, and the results suggest that the student doesn't need any other training signal to avoid the invalid states. This reward structure has been used previously in a single teacher-student setting like ASP [6, 9].

Loss function: To enable student learning, we incorporate a behavioural cloning loss ($\mathcal{L}_{BC}$) for the student, in addition to the actor and critic loss functions ($\mathcal{L}_{RL}$) used in TD3 [3]. Formally, the student agent's loss function $\mathcal{L}(\theta_S) = (1-\delta)\mathcal{L}_{RL} + \delta\mathcal{L}_{BC}$ where δ is a parameter which controls the trade-off between the actor-critic loss and behaviour cloning loss. The actor-critic loss $\mathcal{L}_{RL}$ can be any underlying actor-critic algorithm's (such as TD3) loss function.

The behavior cloning loss is defined as below:

$$\mathcal{L}_{BC} = \mathbb{E}_{(s_t,g_t)\sim D_S}\left[\|\pi_S(a \mid s_t, g_t) - \pi_T(a \mid s_t)\|^2\right]$$

where D_S refers to the student's minibatch sampled during training.

3 EXPERIMENTAL RESULTS

3.1 Experimental settings

We tested our hypothesis on the Fetch-Reach environment and a custom driving simulator. We compare our approach against the most relevant work by OpenAI et al. [6] denoted as ASP in our plots. In our results, the algorithm corresponding to number of teachers as 1 refers to this baseline. Each curve in the plot is obtained by running over 5 different trials with varying seeds and plotting their mean and standard deviation. To make sure that the students with different number of teachers see the same number of goals in one episode, we adjust the multiplier m accordingly (for example, for 16 teachers, we keep the multiplier 1, and for 1 teacher, we keep the multiplier 16).

3.2 Results and discussion

Figure 1(a) show the performance of the student agent in terms of its ability to precisely reach a set of random goals in fetch reach highlighting the importance of multiple teachers to better generalize to random goals. Further, we divide the state space into $5 \times 5 \times 5$ equal parts and measure the number of parts covered by the goals generated so far. Our results show that the state-space coverage increases with an increasing number of teachers as shown in 1(b).

Future work includes investigating the effect of using an explicit diversity component in teacher objectives and trying different diversity metrics. We would also like to decipher what these different teachers learn, for example in a robotics environment, is it possible to understand if any of the teachers propose goals specific to a sub-task?

ACKNOWLEDGMENTS

Part of this work has taken place in the Intelligent Robot Learning (IRL) Lab at the University of Alberta, which is supported in part by research grants from the Alberta Machine Intelligence Institute

(Amii); a Canada CIFAR AI Chair, Amii; Compute Canada; Huawei; Mitacs; and NSERC.

REFERENCES

[1] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems* 29 (2016), 1471–1479.

[2] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2019. Exploration by random network distillation. In *International Conference on Learning Representations.* https://openreview.net/forum?id=H1lJJnR5Ym

[3] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1582–1591. http://proceedings.mlr.press/v80/fujimoto18a.html

[4] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2016. Vime: Variational information maximizing exploration. *Advances in neural information processing systems* 29 (2016).

[5] Richard Meyes, Hasan Tercan, Simon Roggendorf, Thomas Thiele, Christian Büscher, Markus Obdenbusch, Christian Brecher, Sabina Jeschke, and Tobias Meisen. 2017. Motion Planning for Industrial Robots using Reinforcement Learning. *Procedia CIRP* 63 (12 2017), 107–112. https://doi.org/10.1016/j.procir.2017.03.095

[6] OpenAI OpenAI, Matthias Plappert, Raul Sampedro, Tao Xu, Ilge Akkaya, Vineet Kosaraju, Peter Welinder, Ruben D'Sa, Arthur Petron, Henrique P. d. O. Pinto, Alex Paino, Hyeonwoo Noh, Lilian Weng, Qiming Yuan, Casey Chu, and Wojciech Zaremba. 2021. Asymmetric self-play for automatic goal discovery in robotic manipulation. arXiv:2101.04882 [cs.LG]

[7] Silviu Pitis, Harris Chan, Stephen Zhao, Bradly Stadie, and Jimmy Ba. 2020. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In *International Conference on Machine Learning*. PMLR, 7750–7761.

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

[9] Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. 2018. Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.*