Generative Adversarial Skill Estimation in Opponent Modeling

Anonymous authors
Paper under double-blind review

Abstract

In opponent modeling, data about failed actions and models of opponent capabilities can be mined to improve estimates of the strategy gradient and the reliability and stochasticity of certain actions for the given opponent. We want our opponent modeling systems to reason similarly not only about what the opponent plans to do, but also about their probability of success should they choose a given action. The problem of skill estimation (Archibald and Nieves-Rivera, 2018) is closely related, and Bayesian techniques have been proposed for simulated, real-valued games including darts and billiards (Archibald and Nieves-Rivera, 2019). In this paper, I propose a novel method called generative adversarial skill estimation (GASE) to encourage the estimation and the probability of success in RL opponent modeling via introducing an intrinsic reward output from a foundation model generative adversarial network, where the generator provides fake samples of the opponent's actions that help discriminator to identify those failed actions with their probability of success. Thus the agent can identify failed actions that the discriminator is less confident to judge as successful. This work is mainly motivated by the question: How can FMs be used for skill discovery?

1 Introduction

Many real-world physical domains require agents to both plan and execute continuous actions. These planned actions can generally not be executed with perfect precision, resulting in some amount of execution error which will vary from agent to agent. Robust agents in these domains need to plan actions taking into account their execution noise. In different settings, an important characteristic of an agent is the probability distribution over action execution errors that describe their ability to precisely execute actions.

This position paper focuses on the problem of estimating this opponent agent property given observations of the same agent acting in a continuous domain. This problem was first introduced in (Archibald and Nieves-Rivera 2018). The central assumption in that work was that the agent utilized a perfectly rational planning component, which is similar to assumptions made in work on plan recognition in continuous domains (Kaminka, Vered, and Agmon 2018). Thus, introduce the Generative Adversarial Skill Estimation (GASE) in the opponent modeling framework, which aims to study how can FMs be used for skill discovery.

2 The GASE Architecture

Given a policy π, according to (Sham Kakade and John Langford. 2002), the normalized discounted frequency is written as

$$\rho_p(s) = (1-\gamma)\Sigma_{t=0\ldots\infty}\gamma^t P(s_t=s|\pi,s_0)(1)$$

where actions are selected following the policy π and the state transits accordingly. To estimate the failed action, we build negative samples that follow the distribution $g(s)$, and then the probability to be a failed action for state s is given by

$P(1(sisnovel)|s) = 1P(1(sisvisited)|s)$

$$= 1\frac{\rho_p(s)}{\rho_p(s) + g(s)}(2)$$

Therefore, we train a discriminator to represent the probability $D_{\theta d}(s) = \rho_p(s)/\rho_p(s) + g(s)$ indicating the set of opponent's actions . Additionally, we also train a generator $G_{\theta G}$ in order to learn the probability $G_{\theta G} = g(s)$ representing the distribution of negative samples.

In GASE, the generator G is fed with random noise, aiming to generate states as real as they are sampled from the policy interacting with the environment $\rho_p(s)$. The discriminator D aims to discriminate between the successful sampled from $\rho_p(s)$ and the failed actions that are generated from G. As GASE learns, once the agent encounters a novel state with a novel action s_{t+1}+1 after taking some action $a(t)$, D will regard it as a failed action with a low P (1(s is visited)|s), and a large bonus intrinsic reward r_t^i will be assigned to the novel s_{t+1}. Note that if D learns in an online manner where real states are sampled sequentially from the environment, GASE will risk being affected by the high correlation of recent states, and forget the states it has seen before. Thus it is practically effective to employ an additional experience replay buffer M. Moreover, it is worth noting that we take the following techniques within GAEX:

State Abstraction. To efficiently distinguish the successful and the failed actions, we employ sctions abstraction to reduce the action space, since the sensory inputs involve too many useless details for distribution modeling. Thus, we transform the original input into a compact feature space ϕ that keeps the important and ignores the rest.

Choice of Intrinsic Reward Function f. Since the probability of a failed action given by $D_(a)$ will increase as similar states have been visited many times, the intrinsic reward should be a monotonic decreasing function of that probability. Here β is an adjustable hyperparameter. Other function forms are also open to explore.

References

Archibald, C., and Nieves-Rivera, D. (2018). Execution skill estimation. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pp. 18591861.

Archibald, C., and Nieves-Rivera, D. (2019). Bayesian execution skill estimation. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Vol. 33, pp. 60146021.

Kaminka, G. A.; Vered, M.; and Agmon, N. 2018. Plan recognition in continuous domains. In Proceedings of 32nd AAAI Conference.

Sham Kakade and John Langford. 2002. Approximately optimal approximate reinforcement learning. In ICML, Vol. 2. 267–274.