Idea: Online Opponent Modeling with Foundation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Opponent modeling (OM) is the ability to use prior knowledge and observations in order to predict the behavior of an opponent. On the other hand, there has been tremendous research at the intersection of foundation models (FM) and decision-making which holds tremendous promise for creating powerful new systems that can interact effectively across a diverse range of applications. This paper examines the integration of foundation models with opponent modeling and tackles one of the open problems in FMs for decision-making (i) leveraging and collecting decision-making datasets D_{RL}; specifically datasets for the opponent modeling systems in the large-scale human demonstration, which is hard to scale., and (ii) proposing a new framework for opponent modeling: Using FMs as a guiding tool that enhances the agent capabilities in prediction. The goal is to train a policy from a given environment without reward signals. I propose using foundation models (FMs), i.e., large language models (LLMs) and vision-language models (VLMs), to achieve this goal. The LLM generates instructions that help the agent to learn features of the behavior of the opponent and ultimately enables the agent to exploit the opponent's strategy in the current environment $d(s_0)$. In contrast, the VLM works as a policy-guided learning. The internet-scale knowledge capacity of recent FMs enables automating impractical human effort in the RL framework [1]. Existing works query pre-trained LLMs for tasks to learn [2], language-level plans [3], and language labels [4]; or use pre-trained VLMs to obtain visual feedback [5]. ELLM [6] uses LLMs to propose new tasks for agents to learn. A line of work [7] specifically focuses on using FMs for the Minecraft domain, while none of the works integrate pre-trained LLM and VLM for opponent modeling. Inspired by [8], this work is mainly motivated by two questions: How to leverage and construct datasets for decision-making D_{RL} i.e. FMs and OM? And can we teach RL agents to predict opponents' actions and strategies accurately in opponent modeling environments without human supervision?

1 Introduction

In a Partially-Observable Stochastic Game (POSG) [9] for a basic formalization of the competitive environment. A POSG is defined by a tuple $<I, S, {O^i}_i, A, T, R^i i, \Omega^i i>$, where I = 1, 2, . . . , N is the set of agents. S is the state space. O^i is the observation space of agent i. $A = A^1 \times A^2 \times \ \times A^N$ is the joint action space. $T : S \times A \times S \rightarrow [0, 1]$ denotes the transition dynamics, which defines the probability distribution on the next state given the previous state and the joint action. $R_i : S \times A \times S \rightarrow$ R denotes the reward function of agent i. $\Omega^i : S \times A \times O^i \rightarrow [0, 1]$ denotes the agent i's observation function, which defines the probability distribution over its possible next observation given the previous state and the joint action.

Submitted to 38th Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024). Do not distribute.

In this study, I utilize 1 to denote the controlled agent and 1 to denote the opponent and focus on modeling one opponent. I assume that the opponent's policy originates from a set of fixed policies $\Pi = \{\pi^{1,k}(a^1|o^1)\}_{k=1,2,\dots,k}$, which are obtained by the scripts or RL algorithms pre-training.

1.1 Notations

The first step is to generate a set of imagined task instructions that are useful for learning behaviors. Given the proposed set of N-numbers of task instructions $\{\delta^{(}i)\}_{i=1\dots N}$ and their corresponding initial states, our goal is to train a multi-task policy $\pi(a|s,\delta)$ that follows the instructions. To accomplish this, VLM can be used as a policy-guided learning, which trains a multi-task policy in the training environment using the obtained instructions. The policy is trained to follow the given instruction by maximizing the VLM "alignment score" between the current observation and the instruction as its reward. Specifically, the reward is defined by:

$$r_t = r(o_{tH:t}, \delta) = \frac{\phi_v(o_{tH:t})^T \phi_T(\delta)}{|\phi_v(o_{tH:t})| \cdot |\phi_T(\delta)|} \tag{1}$$

where o_t is the visual observation of time step t with $o_{tH:t}$ implying the sequence of observations with size of H, $|\cdot|$ refers to L2-norm of a vector, ϕ_T and ϕ_v are the text and video encoder of the VLM, δ is the language instruction, and H is the length of video that the VLM takes.

$$\Sigma_{i=1\dots N}\, E_o t_{\sim}\pi, \rho, P[\Sigma_t \hat{r}(o_{tH:t})\delta_i)] \tag{2}$$

.

2 Opponent Modeling

Assuming that the interaction between the controlled agent and the opponent policy $\pi^{^{1},^{k}}$ generates their respective trajectories, denoted as $\tau^1,^k = (o_0^{1,k}, a_0^{1,k}, r_0^{1,k}, o_1^{1,k}, a_1^{1,k}, r_1^{1,k}, \dots) \in^{-\ 1},^k$ and $\tau^{^{1,k}=(o_0^{-1,k},a_0^{-1,k},r_0^{-1,k},o_1^{-1,k},a_1^{-1,k},r_1^{-1,k},\dots)\in\tau^{-1-1k}}$. The resultant dataset is thus denoted as $D^k = \tau^{1,k}, \tau^{^{1,k}}$. Within the context of offline learning, we presume the availability of the dataset $D^{off} = D^k{}_{k=1,2,\dot{K}}$. Specifically, $\tau^{1,k}$ is acquired through interactions with $\pi^{-1,k}$ while employing its approximate best response policy $\pi^{1,k,*}$, usually with certain noise.

The objective of OM is to use D^{off} to pre-train an opponent-aware adaptive controlled agent policy $M_\theta(a^1|o^1; D)$ and deploy M into a new environment with an unknown test opponent policy set $\Pi^t est$t, such that the controlled agent achieves the maximum expected return (i.e., cumulative reward):

$$max E_\pi^{-1} \sim \Pi^t est, D^{off}, T, \Omega[\Sigma_{t=0\dots\infty}\, R_t^1 | a_t^1 \sim M_\theta.\pi^{-1}] \tag{3}$$

D is the opponent's information data, sampled from Doff during offline pre-training and must be collected during deployment.

3 How to Leverage or Collect Datasets

One key challenge in applying foundation models to decision-making lies in the dataset gap: the broad datasets from vision and language D and the task-specific interactive datasets D_{RL} can be of distinct modalities and structures. For instance, when D consists of videos, it generally does not contain explicit action labels indicating the cause-effect relationship between different frames, nor does it contain explicit reward labels indicating which videos are better than others, whereas actions and rewards are key components of D_{RL}. Despite this gap, broad video and text data can be made more task-specific through post-processing ($D \to D_{RL}$), leveraging hindsight relabeling of actions and rewards (e.g., using human feedback). Meanwhile, decision-making datasets can be made more broad and general ($D \to D_{RL}$) by combining a wide range of tasks-specific datasets (e.g., Gato). Below we provide a list of examples of D and D_{RL} that can be used for research in foundation models for decision-making, and propose additional approaches for bridging the gap between D and D_{RL}. In the manuscript of [10] proposed bridging D and D_{RL}. To enable better datasets tailored for decision-making, one can either increase the scale of D_{RL}. by large-scale logging and merging task-specific sets of interactive data or by relabeling D with action and reward information. One could also consider augmenting D_{RL}. with metadata, such as informational and instructional texts and videos.

References

[1] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In arXiv preprint arXiv:2307.15818, 2023

[2] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language models. International Conference on Machine Learning, 2023.

[3] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 11523–11530, 2023. doi: 10.1109/ICRA48891.2023.10161317.

[4] Jesse Zhang, Jiahui Zhang, Karl Pertsch, Ziyi Liu, Xiang Ren, Minsuk Chang, Shao-Hua Sun, and Joseph J Lim. Bootstrap your own skills: Learning to solve new tasks with large language model guidance. In 7th Annual Conference on Robot Learning, 2023. URL https: //openreview.net/forum?id=a0mFRgadGO.

[5] Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar, and Aravind Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? Learning for Dynamics and Control Conference, 2022.

[6] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language models. International Conference on Machine Learning, 2023.

[7] Yevgen Chebotar, Quan Vuong, Karol Hausman, Fei Xia, Yao Lu, Alex Irpan, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, Keerthana Gopalakrishnan, Julian Ibarz, Ofir Nachum, Sumedh Anand Sontakke, Grecia Salazar, Huong T Tran, Jodilyn Peralta, Clayton Tan, Deeksha Manjunath, Jaspiar Singh, Brianna Zitkovich, Tomas Jackson, Kanishka Rao, Chelsea Finn, and Sergey Levine. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. In 7th Annual Conference on Robot Learning, 2023. URL https://openreview.net/forum?id=0I3su3mkuL.

[8] Taewook Nam, Juyong Lee, Jesse Zhang, Sung Ju Hwang, Joseph J. Lim, Karl Pertsch. LiFT: Unsupervised Reinforcement Learning with Foundation Models as Teachers. 2023. arXiv:2312.08958v1

[9] Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. 2023. arXiv preprint arXiv:2011.00583, 2020

[10] Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, Dale Schuurmans. Foundation Models for Decision Making: Problems, Methods, and Opportunities. arXiv:2303.04129v1