Goal Representations for Goal-Oriented Agents in Reinforcement Learning

Weinan Qian
Yuanpei College
Peking University
ypqwn@stu.pku.edu.cn

Abstract

Humans possess inherent capabilities to act in alignment with their intentions and showcase an early ability to discern the intentions of others through observation. Harnessing inspiration from this innate quality, the realm of Reinforcement Learning (RL) places significant importance on delving into meaningful goal representations within a computational framework. In the intricate task of designing an agent responsible for observing the environment and making decisions to achieve specific objectives, it becomes imperative to engage in the discourse regarding the apt representations of goals and the potential pathways to actualize these objectives. We introduce two conceivable approaches for representing goals in goal-oriented agents: reward-based goal representations and subtask-based goal representations. Each method of representations comes with its own set of merits and limitations.

1 Introduction

We inhabit in a world teeming with animate, goal-directed agents, whose agency encompasses the capacity to perceive, plan, make decisions and achieve their goals [14]. Humans, as a distinctive example of animate agents, primarily undertake actions to fulfill their intentions arising from their beliefs and desires [1]. Unlike other creatures, such as the New Caledonian crow that bends twigs into hooks to better reach its prey and satisfy its "intent" [9, 13], humans possess the unique ability to elucidate and foresee others' actions by attributing causal intentional mental states to them [4]. A plethora of research in the field of Theory of Mind (ToM) provides compelling evidence that young children, by the age of 5, adeptly apply a relatively sophisticated mentalistic interpretational strategy to explain and predict the behavior of other agents [5].

These research sheds light on the significance of intentionality in the context of Reinforcement Learning (RL), which addresses the challenge faced by an agent learning behavior through trial-and-error interactions with a dynamic environment [7]. Naturally, researchers in this domain grapple with the task of enabling agents to achieve specific goals in dynamic environments. To tackle this fundamental challenge, one must contemplate a pivotal question: how can we effectively represent the goals of an agent? Without a clear representation of the goal, the agent risks becoming an "opaque" or "black box" agent, rendering its actions devoid of interpretability [10]. Designing an appropriate method to represent the goals of an agent stands as the initial step in RL, fostering actions that are more directional and purposeful.

In this paper, we delve into two distinct approaches for representing goals in goal-oriented agents. In Sec. 2, we introduce the reward-based goal representation, a classical method in RL that incentivizes agents to take actions aimed at maximizing their reward function and, consequently, moving closer to ideal states. Sec. 3 outlines a more systematic method involving the decomposition of goals into several subtasks. Agents then strategically accomplish these subtasks in a planned sequence to ultimately achieve the overarching goal. In addition to presenting the fundamental definitions and examples of these methods, each section conducts a brief analysis of their respective strengths and weaknesses.

2 Reward-Based Goal Representations

Typically, an agent learns to make decisions through interactions with an environment to maximize its cumulative reward. Upon reaching a specific state, indicative of achieving pre-defined goals, the agent receives a positive reward, encouraging a continuation of similar actions. Formally, we often employ a Markov decision process (MDP) specified by a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ to characterize the environment faced by the agent. Here, $\mathcal{S}$ denotes the set of states, $\mathcal{A}$ represents the set of actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ signifies the transition function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ indicates the reward function, and γ is the discount coefficient. In this environment, the agent strives to identify the optimal stochastic policy $\pi_\star : \mathcal{S} \to \Delta\mathcal{A}$ ($\Delta\mathcal{A}$ signifies the simplex of $\mathcal{A}$, *i.e.*, all possible probabilistic distributions over the set of actions $\mathcal{A}$) such that $q_{\pi_\star}(s, a) = q_\star(s, a)$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, where $q_\pi(s, a)$ is the expected reward given policy π with initial state s and action a, and $q_\star = \text{argmax}_\pi\, q_\star(s, a)$. This searching process oriented by the reward leads us to define this method of goal representations as reward-based goal representations.

Previous research has widely adopted such representations, exemplified by Kolobov et al. [8] extending MDP to Stochastic Shortest Path (SSP) MDP. They introduced new classes of MDP to effectively address the shortest path planning problem, even permitting the occurrence of dead ends under progressively weaker assumptions. Taking a step further, Fan et al. [3] delved into task-agnostic reward representations in the context of *Minecraft*, where agents operate in an open-ended environment and should have the capacity to solve tasks to fulfill tremendous or even infinite varieties of goals. Their contribution, MINECLIP is a contrastive video-language model that learns to correlate video snippets and natural language descriptions (See Fig. 1). To elaborate further, they train a dense reward function $\Phi_\mathcal{R} : G \times V \to \mathbb{R}$ which yields high rewards when the behavior depicted in the video faithfully follows the language description, where $g \in G$ represents a descriptive goal in language form, and $v \in V$ is a video snippet. This is achieved by utilizing collected *Minecraft* video snippets with time-aligned English transcripts and optimizing InfoNCE objective [2, 6]. The learned reward function is then leveraged to train a language-conditioned policy network, taking raw pixels as input and predicting discrete control in *Minecraft*.

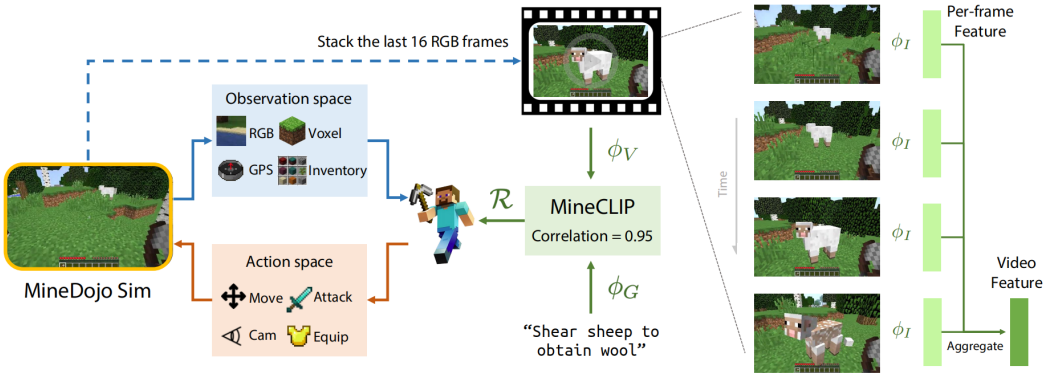

Figure 1: The framework of MINECLIP [3].

Reward-based goal representations are evidently intuitive and easily comprehensible. They offer a direct criterion for agents to refernce when taking actions, and make it convenient for researchers to implement such reward-based method in a computational framework. This method also allow agents to iteratively adjust their actions through optimization methods. However, tailoring reward-based representations to fulfill specific goals often necessitates nuanced strategies, aand their efficacy is typically confined to a limited set of tasks, thereby lacking generalization capacity. Additionally, a noteworthy drawback arises from the sparsity of numerous reward functions, *i.e.*, they assign a value of 0 to most inputs $(s, a) \in \mathcal{S} \times \mathcal{A}$. This sparsity can impede the convergence of the proposed RL model, leading to prohibitively slow convergence rates.

3 Subtask-Based Goal Representations

At times, goals can be exceedingly intricate, posing a formidable challenge in the creation of a straightforward reward function for agents to achieve such complexity. Recognizing the inherent

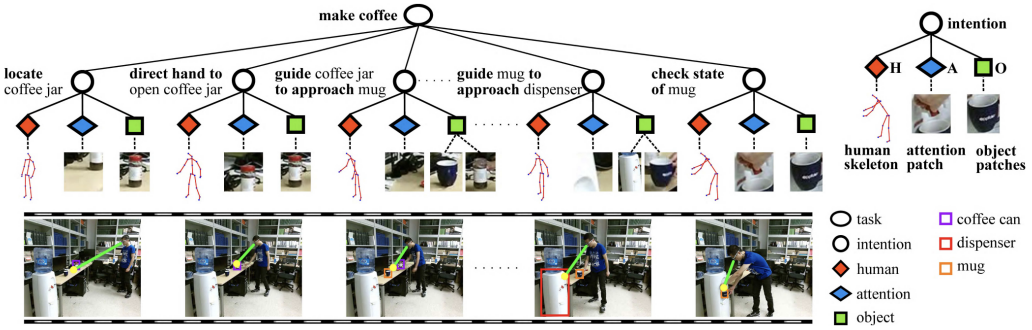

Figure 2: A task is modeled as sequential intentions in terms of hand-eye coordination with a human-attention-object (HAO) graph [12].

difficulty in satisfying these intricate goals, subtask-based goal representations emerge as a remedy, enabling agents to systematically attain complex goals step by step. This method involves representing the original goal as a series of subtasks that the agent must complete in a specific order. Consequently, it effectively decomposes the overarching task, transforming the pursuit of a complex goal into a sequence of more manageable components.

Numerous emerging research efforts are dedicated to exploring effective subtask-based goal representations. Wei et al. [12] introduced a hierarchical model of HAO, unifying tasks, intentions, and attention within a comprehensive framework. In this model, more complex tasks are represented as sequential intentions transitioning to another. As an example, the task *make coffee* is delineated into 8 sequential intentions, each more fundamental and easier to accomplish (See Fig. 2). In the realm of RL applied to *Minecraft*, where intricate goals like obtaining a *diamond* abound, Wang et al. [11] presented the "Describe, Explain, Plan and Select" (DEPS) framework (See Fig. 3). In this framework, a planner initially receives a language instruction, devises several possible subtask paths, and submits them to a selector to choose the most efficient path for agent execution. After execution, a controller provides feedbacks to a descriptor that narrates the current state. Subsequently, a explainer generates potential issues based on the descriptor's descriptions and presents these explanation to the planner for re-planning subtask paths. The planning process terminates either when the agent successfully achieves the goal or when the number of planning steps reaches a specified threshold.

The advantages of subtask-based goal representations are apparent. They offer the capacity to represent intricate goals more effectively and enabling agents to better attain these goals, coupled with enhanced interpretability. However, this method is not without imperfections and comes with

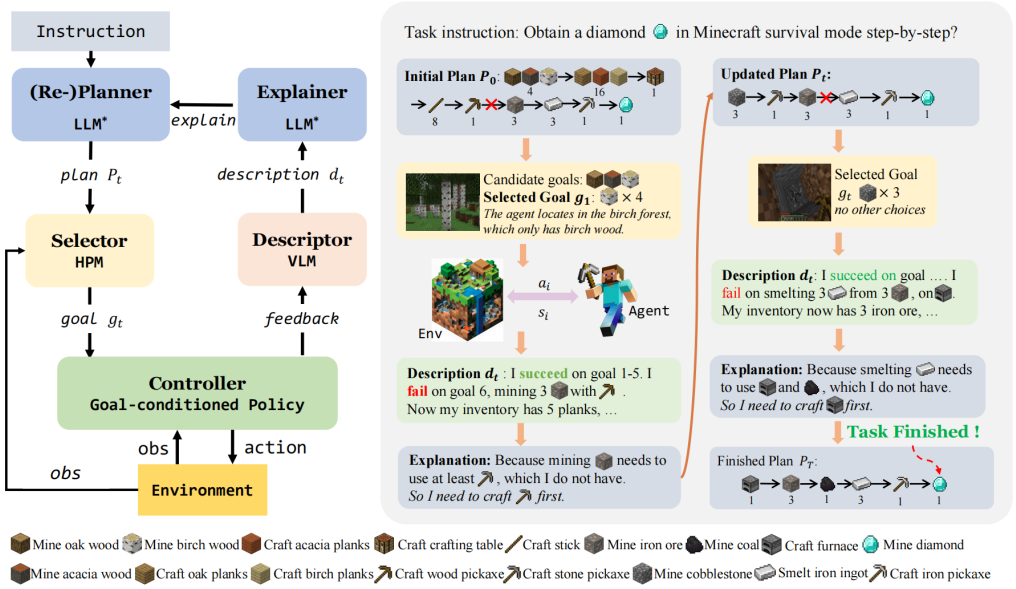

Figure 3: Overview of the proposed DEPS architecture [11].

certain limitations. Notably, the intricacy of the subtask space poses challenges in constructing a high-quality and fine-grained subtask space, essential for the agent's effective decomposition of goals. Coordinating the execution of subtasks and navigating the complexity of their sequential interactions may also prove challenging, demanding that the agent learns effective policies for both selecting and sequencing subtasks.

4 Conclusion

In this paper, we present two distinct methods for representing goals in an agent: reward-based goal representations and subtask-based goal representations. On one hand, the former is intuitive and classical, yet it is constrained to handling a limited number of simpler goals. On the other hand, the latter exhibits the ability to represent a broader spectrum of goals, but challenges arise from the complexity of the subtask space and the intricacy of determining the execution sequence for subtasks. Each method carries its own set of benefits and defects, prompting the necessity to carefully choose the most suitable approach based on specific research requirements and the characteristics of the actual environments under consideration.

References

[1] Dare A. Baldwin and Jodie A. Baird. Discerning intentions in dynamic human action. *Trends in Cognitive Sciences*, 5(4):171–178, 2001. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020. 2

[3] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[4] György Gergely and Gergely Csibra. Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, 7(7):287–292, 2003. 1

[5] György Gergely, Zoltán Nádasdy, Gergely Csibra, and Szilvia Bíró. Taking the intentional stance at 12 months of age. *Cognition*, 56(2):165–193, 1995. 1

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[7] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996. 1

[8] Andrey Kolobov, Mausam, and Daniel S. Weld. A theory of goal-oriented mdps with dead ends. In *arXiv preprint arXiv*, 2012. 2

[9] Douglas Main. Like chess players, these crows can plan several steps ahead, 2019. URL `https://www.nationalgeographic.com/animals/article/new-caledonian-crows-plan-ahead-with-tools`. 1

[10] Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33:673–705, 2019. 1

[11] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. In *arXiv preprint arXiv*, 2023. 3

[12] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[13] Nat Geo WILD. Tool-making crows are even smarter than we thought, 2018. URL https://www.youtube.com/watch?v=UZM9GpLXepU&t=56s. 1

[14] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 1