Learn Goal Representations for Goal-Conditioned Reinforcement Learning

Zhancun Mu
YuanPei College
Peking University
yhbylch@stu.pku.edu.cn

Abstract

Goal-conditioned reinforcement learning (GCRL) aims to design agents capable of solving multiple tasks. However, choosing, representing, and learning goals remains an open area of research. This essay first reviews goal representation and learning methods in GCRL. It then discusses challenges and future directions.

1 Introduction

One of humans' most remarkable behaviors is acting guided by diverse goals [11]. Goals can be abstract or detailed, making unified representation challenging. This essay focuses mainly on using goals in reinforcement learning (RL).

Unlike standard RL, which aims to learn a policy solely based on the current state, GCRL requires the agent to make decisions per different goals. One approach is specifying reward functions for each goal. However, this can be difficult to design and tune, especially with a large goal space. Thus, agents must learn goal representations. This essay first formulates the GCRL problem. It then reviews existing methods for goal representation and learning. Finally, it discusses challenges and future directions.

2 Preliminaries

This section introduces the RL and GCRL problem formulation briefly.

Standard RL tasks can be modelled as Markov Decision Process (MDP), denoted as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action space respectively, $\mathcal{P}$ is the transition probability, $\mathcal{R}$ is the reward function, and γ is the discount factor. The problem requires agent to learn a policy $\pi(a|s)$ to maximize the expected total reward

$$J(\pi) = \mathbb{E}_{a_t \sim \pi, s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)} \left[\sum_t \gamma^t r(s_t, a_t) \right].$$

Though decomposing long-horizon tasks into subgoals is also a part of GCRL, there we mainly focus on develop a policy with a given goal. The formulation augments the MDP with $(\mathcal{G}, p_g, \phi)$, where $\mathcal{G}$ is the goal space, p_g is the goal distribution, and $\phi : \mathcal{S} \to \mathcal{G}$ is a mapping function. The reward function $\mathcal{R}$ is related to the goal, i.e. $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to \mathbb{R}$, and the policy $\pi(a|s, g)$ is goal conditioned.

3 Goal Representation

3.1 Definition of Goal

First, the definition of a goal in RL literature should be understood. Some common definitions [11] are:

- **Desired Goal**: A desired goal is a required task for the agent to complete. It can be either provided by the environment or generated internally.
- **Achieved Goal**: An achieved goal is the goal state, i.e. $g = \phi(s)$, that the agent has reached.
- **Behavioral Goal**: A behavioral goal is a goal that the agent is currently pursuing.

For policies without subgoal decomposition, the desired goal and behavioral goal are the same.

3.2 Goals in GCRL

This section introduces the types of goals used in GCRL and their representations.

3.2.1 State as Goals

Representation of future state is a common goal used in GCRL. The goal can be the target state, or an imagined sequence of actions reaching the target. In many settings, $\mathcal{S}$ and $\mathcal{G}$ are identical, meaning ϕ is the identity function. For example, [1] develops a policy directly conditioned on a future state. However, due to high image dimensionality, latent vector representations are useful. One powerful tool is the CLIP model [14], which bridges the gap between text and image. Fan et al. [5] propose MineClip, pretrained on MineDojo's massive YouTube videos. To leverage pretrained image embeddings, Lifshitz et al. [10] finetunes a Video Pretraining (VPT) model conditioned on these embeddings. Variational autoencoders (VAEs) [8] also map states to goals, as in [7].

3.2.2 Language as Goals

Despite CLIP's language use, directly conditioning RL on language is also an active research direction. Standard RL and imitation learning lack prior knowledge and generalization. However, language encodes abstract meaning and knowledge for generalization, communication, and conveying intentions. Recent representation learning advances vectorize language, enabling language integration in RL. Using language as a goal in RL is now an active research area.

One approach interleaves language reasoning and action in a unified policy. Modeling decision-making as sequence generation is widely studied due to Transformer success. Decision Transformer (DT) [4] advances this direction. To leverage language and DT, Mezghani et al. [12] augment a Transformer policy with word outputs, as in Fig. 1.

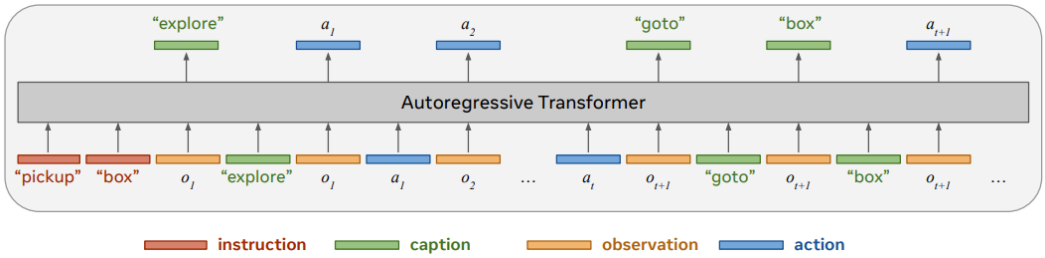

Figure 1: Given an instruction, the transformer policy can generate language based reasoning tokens interleaved with sequence of actions in the environment. Image and caption credit: [12].

3.2.3 Reward as Goals

Obtaining a reward function is an advantage of using goals in GCRL. For instance, the distance between the current and goal states can provide intrinsic rewards. Alternatively, the desired return can serve as the goal. Kumar et al. [9] propose learning a policy $\pi(a|s, R)$ where R is the estimated return. Similarly, DT uses return-to-go (RTG) to guide the policy. This idea partially addresses suboptimal offline data.

4 Extracting Goals from Data

GCRL lacks data with appropriate goal annotations. YouTube videos offer abundant MINECRAFT tasks, but annotating them with text goals is challenging. Fan et al. [5] search phrases like "MineCraft

Tutorial" and relabel with human annotation. However, this approach is expensive and limited to language goals. This introduces methods to extract goals from data.

4.1 Goal Relabeling

Hindsight Experience Replay (HER) inspires labeling future states as goals. For instance, Lifshitz et al. [10] randomly select episode timesteps and hindsight relabel intermediate goals as visual MINECLIP embeddings (Fig. 2). Additionally, hidden training states can serve as goals. Pertsch et al. [13] first train an LSTM policy, then use encoder embeddings as goals. Key video frames also extract goals. [1] cluster video frames, using cluster centers as goals. They first use future video frames as goals to train a goal-conditioned policy and then use the clustered goals for fine-tuning.

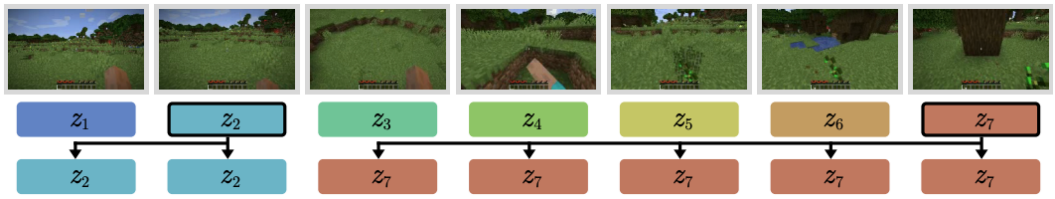

Figure 2: Relabeling strategy in STEVE-1. Image credit: [10].

4.2 Inverse Reinforcement Learning

Inverse reinforcement learning (IRL) also learns goals from demonstrations by inferring reward functions. This aims to model expert preferences. Classical IRL methods include feature matching and maximum entropy, as reviewed in [2]. Humans use inverse planning to understand others' intents via beliefs and desires. Baker et al. [3] formalize this as Bayesian inverse planning, i.e.

$$P(Goal|Actions, Environment) \propto P(Actions|Goal, Environment)P(Goal),$$

though this is only a preliminary framework requiring further development.

5 Challenges and Future Directions

GCRL has challenges including generalization and sample efficiency. However, some phenomena are not well addressed yet. Humans can self-generate goals through exploration, but current methods struggle transferring this to complex environments like MINEDOJO [5]. Some methods intrinsically learn skills by maximizing mutual information between skills and behaviors, as in [15]. However, these cannot transfer well to complex environments.

Injecting prior knowledge is also challenging. Humans infer social relationships from 2D motion in the Heider-Simmel animation [6], using relationship knowledge to understand movements. We also view others' actions as rational and goal-directed. Learning, representing, and appropriately incorporating such priors in systems remains an open problem requiring better understanding of human cognition.

References

[1] Anonymous. Pre-training goal-based models for sample-efficient reinforcement learning. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=o2IEmeLL9r`. under review. 2, 3

[2] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021. 3

[3] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009. 3

[4] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021. 2

[5] 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 *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=rc8o_j8I8PX. 2, 3

[6] Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944.

[7] Riashat Islam, Hongyu Zang, Anirudh Goyal, Alex Lamb, Kenji Kawaguchi, Xin Li, Romain Laroche, Yoshua Bengio, and Remi Tachet Des Combes. Discrete factorial representations as an abstraction for goal conditioned reinforcement learning. *arXiv preprint arXiv:2211.00247*, 2022. 2

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[9] Aviral Kumar, Xue Bin Peng, and Sergey Levine. Reward-conditioned policies. *arXiv preprint arXiv:1912.13465*, 2019. 2

[10] Shalev Lifshitz, Keiran Paster, Harris Chan, Jimmy Ba, and Sheila McIlraith. Steve-1: A generative model for text-to-behavior in minecraft. *arXiv preprint arXiv:2306.00937*, 2023. 2, 3

[11] Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. *arXiv preprint arXiv:2201.08299*, 2022. 1

[12] Lina Mezghani, Piotr Bojanowski, Karteek Alahari, and Sainbayar Sukhbaatar. Think before you act: Unified policy for interleaving language reasoning with actions. *arXiv preprint arXiv:2304.11063*, 2023. 2

[13] Karl Pertsch, Youngwoon Lee, and Joseph Lim. Accelerating reinforcement learning with learned skill priors. In *Conference on robot learning*, pages 188–204. PMLR, 2021. 3

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[15] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019. 3