HILONet: Hierarchical Imitation Learning from Non-Aligned Observations

Shanqi Liu[1],Junjie Cao[1],Wenzhou Chen[1],Licheng Wen[1],Yong Liu[1]
1. The State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang, 310027, China
E-mail: shanqiliu@zju.edu.cn

Abstract: It is challenging learning from demonstrated observation-only trajectories in a non-time-aligned environment because most imitation learning methods aim to imitate experts by following the demonstration step-by-step. However, aligned demonstrations are seldom obtainable in real-world scenarios. In this work, we propose a new imitation learning approach called Hierarchical Imitation Learning from Observation(HILONet), which adopts a hierarchical structure to choose feasible sub-goals from demonstrated observations dynamically. Our method can solve all kinds of tasks by achieving these sub-goals, whether it has a single goal position or not. We also present three different ways to increase sample efficiency in the hierarchical structure. We conduct extensive experiments using several environments. The results show the improvement in both performance and learning efficiency.

Key Words: Imitation Learning from Observation, Hierarchical

1 Introduction

Robots can acquire complex behavior skills suitable for various unstructured environments through learning. Two of the most prevalent paradigms for behavior learning in robots are imitation learning(IL) and reinforcement learning(RL). RL methods can theoretically learn behaviors that are optimal with respect to a clear task reward. However, it usually takes millions of training steps to converge. IL methods, on the other hand, can learn faster by mimicking expert demonstrations. But in many real world scenarios, the demonstrations are hard to obtain as we may not able to obtain expert's actions or the expert has a different action space. In such a case, the more-specific problem of imitation learning from observation(ILfO) must be considered.

Previous works in ILfO focus on mimicking an expert skill by following the demonstration step-by-step, such as TCN[1]. They aim to imitate human demonstrations without access to the underlying actions, and they assume that a demonstration can be temporally aligned with the agent's actions. This assumption does not hold when environments do not have a constant number of steps to end, which is common in real world scenarios. As a result, there are a few works forcing on making the non-time-aligned demonstration be time aligned[2]. However, instead of strictly following a demonstration step-by-step, a more natural way is to select the observations that are feasible to achieve adaptively. In this case, the key to ILfO is that how to choose the feasible goals. This task is similar to the goal-based hierarchical reinforcement learning's(HRL) task. However, although few works[3] are drawing on hierarchical reinforcement learning and use it in imitation learning, they did not use it to choose the sub-goal from all demonstration trajectories.

In our work, we propose a novel hierarchical reinforcement learning method that can flexibly choose an observation from the expert's trajectories and use it as the sub-goal to follow. The whole structure of our method consists of two-part, high-level policy and low-level policy. High-level policy outputs the sub-goal chosen from expert trajectory every a few steps and low-level policy takes it as the sub-goal to achieve. We theoretically prove that this structure can solve all two kinds of tasks. One can be effectively described by a single goal observation such as navigation. As well as the other kind of tasks such as swimming, which are usually described by a sequence of key observations rather than a single goal. To the best of our knowledge, most tasks in real world scenario can be classified into these two categories. Thus, our method has broad applicability. Furthermore, to increase the sample efficiency, we propose several methods to overcome the non-stationary in hierarchical structure.

Finally, we test our method and the state-of-art imitation learning methods in five different environments. The result indicates our method outperforms all comparing methods. And we demonstrate that our method can not only reach the goal observation but also mimic the expert behavior as closely as possible.

In summary, the main contributions of our work include:

- We propose a new way of learning from observation using hierarchical reinforcement learning structure to choose feasible sub-goal, which can solve all kinds of non-time-aligned environments.
- We increase the sample efficiency of hierarchical reinforcement learning by overcoming the non-stationary.

2 Method

In our approach, the whole policy is consisted of two part, high policy $\pi_{high}(o_g|o_t;\theta_h)$ and low policy $\pi_{low}(a|o_t,o_g;\theta_l)$, where o_g is sub-goal chosen from expert trajectory observations for low policy, o_t is current observation. We use DDPG[4] algorithm to train both high policy and low policy.

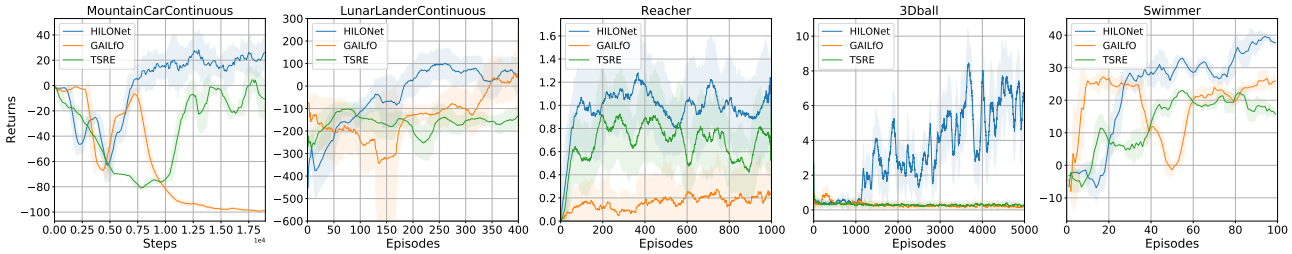

Fig. 1: Comparison of HILONet(ours) to GAILfO and reward engineering baselines.

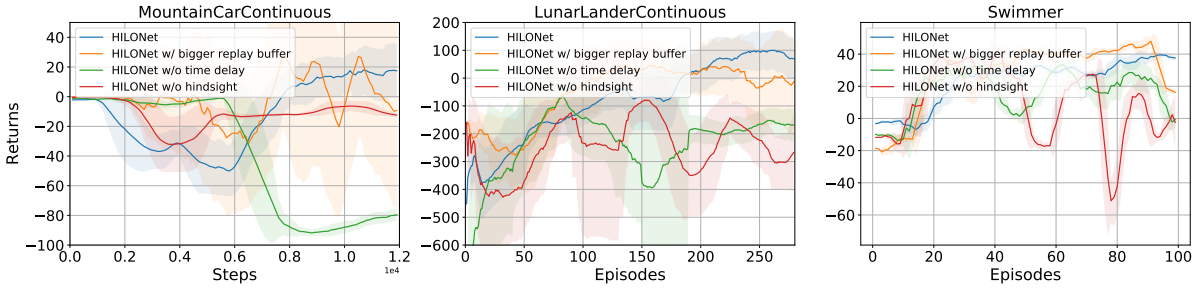

Fig. 2: Ablation experiments

The high policy is charged for choosing a feasible demonstrated observation depending on current observation so low policy is capable of achieving that sub-goal in certain steps. Here, we first define $\tau_i = \{d_1^i, d_2^i, ..., d_T^i\}$ as one trajectory of demonstrations. Then we have $D = \{\tau_1, \tau_2, ..., \tau_T\}$ means all trajectories we used in training process. High policy's action consists of two rates between 0 and 1, the first dimension of action stands for which trajectory in D is chosen, and second action dimension is the index of observation chosen in the trajectory demonstration. In this way, the sub-goal can be formed as:

$$o_g = \pi_{high}(a_h^1, a_h^2 | o_t; \theta_h) = D\{a_h^1, a_h^2\} \tag{1}$$

The low policy focuses on interacting with the environment and find a way to achieve the sub-goal provided by the high policy in certain steps. It takes $\{o_g, o_t\}$ as input and output a_t as inter-action that interacts with the environment. It can be viewed as

$$a_t = \pi_{low}(a_t | o_t, o_g; \theta_l) \tag{2}$$

The rewards for both policy are designed to encourage the policy imitating the demonstration. For low policy, we can use the Euler distance of goal observation and current observation and a sparse reward that is given only when the agent achieves the current sub-goal. We define if $|o_g - o_t| < \epsilon$ then we consider agent has achieved the sub-goal. The overall reward of low policy can be viewed as:

$$r_{low}(o_t, o_g) = \begin{cases} -|o_g - o_t|^2 & \|o_g - o_t\| > \epsilon \\ -|o_g - o_t|^2 + r & \|o_g - o_t\| < \epsilon \end{cases} \tag{3}$$

As for the high policy, we consider it should guide the low policy to accomplish the specific tasks. In this way, we use reward related with which phase agent is right now. This reward can be evaluated by $I(o_g)$ which defined as the index of o_g in its own trajectory. Here, we define $I(o_g) = 0$ if o_t is not in expert trajectory, which can punish agent when it deviates from the expert trajectory. The overall reward of high policy can be viewed as:

$$\Delta a_h^2 = \pi_{high}(o_t) - \pi_{high}(o_{t-\Delta t}) = I(o_g^i) - I(o_g^{i-1}) \tag{4}$$

$$r_{high}(o_g^{i-1}, o_g^i) = \begin{cases} 1 + \alpha \cdot (I(o_g^i) - I(o_g^{i-1})) & \|o_g^i - o_t\| < \epsilon \\ 0 & \|o_g^i - o_t\| > \epsilon \end{cases} \tag{5}$$

This reward can obviously solve these tasks which are described by a single goal position. Additionally, we can theoretically prove that the proposed reward can also solve tasks described by a sequence of key observations.

Furthermore, to increase the sample efficiency, we propose three methods to overcome the non-stationary in hierarchical structure. First, we use a hindsight replacement method to transfer the non-optimal transitions of high policy in HRL into optimal ones. We also propose a time-delay training method to stable the low policy. Additionally, we choose differentiated experience pools for high-level policy and low-level policy.

3 Findings

We find that our method has three advantages. First, comparing to another reward engineering imitation learning from observation methods, our method can plan dynamically thus it can solve the non-time-aligned problem. Furthermore, we prove our method can adopt in all kinds of environments, no matter the tasks are described by a single goal observation or a sequence of key observations. Additionally, our method can use information from multiple trajectories simultaneously, while most reward engineering methods can only imitate one

trajectory if it imitates an expert step-by-step.

Second, our method is based on reward engineering, so comparing with adversarial methods, our method has access to observation information directly, which can offer more dense reward. This means our method do not require as many demonstration examples as adversarial imitation algorithms do to learn an excellent policy.

The third advantage is that using a hierarchical framework policy can naturally divide a complex task into two more straightforward tasks, which will accelerate learning in sequential decision-making tasks.

We prove these advantages by comparing our method to several the state-of-art imitation learning methods in five different environments. We compare our method(HILONet) to other two imitation learning from observation methods, GAILfO[5] and reward engineering baselines(TSRE). Results show test performance over the number of collected episodes or steps. All tests are evaluated over three seeds and using 20 or 30 experts' trajectories in different environments. The result in Fig. 1 indicates our method outperforms all comparing methods. And we demonstrate that our method can not only reach the goal observation but also mimic the expert behavior as closely as possible. Furthermore, we test the effect of these three ways of overcoming non-stationary, as shown in Fig. 2. As a result, we find they improve the effectiveness of our method.

4 Conclusion

In this paper, we introduced a new imitation learning from observation method, hierarchical imitation learning from observation (HILONet), using hierarchical reinforcement structure to choose observations from expert trajectories' observations as goals. By achieving these goals, our method can imitate expert with only observations offered. We give the theoretical proof that our method has the ability to solve tasks with single goal position and tasks described by a sequence of key observations. Additional, we propose three different ways to overcome the non-stationary problem in hierarchical structure to increase sample efficiency. We evaluate the method with extensive experiments based on five different environments, including both these have a single goal position and those do not have a specific target. The result shows that HILONet can solve all kinds of tasks and improves the training procedure of imitation learning. And it outperforms all comparison methods in every environment. Finally, these three ways of overcoming non-stationary are proven to be effective by the result.

References

[1] S. Pierre, Time-Contrastive Networks: Self-Supervised Learning from Video, in *Proceedings - IEEE International Conference on Robotics and Automation*, 2018: 1134–1141.

[2] L. Fangchen, State alignment-based imitation learning, accepted.

[3] L. Youngwoon, To Follow or not to Follow: Selective Imitation Learning from Observations, accepted.

[4] L. Timothy, Continuous control with deep reinforcement learning, in *arXiv preprint arXiv:1509.02971*, 2005.

[5] T. Faraz, Generative adversarial imitation from observation, in *arXiv preprint arXiv:1807.06158*, 2018.