MaMiC: Macro and Micro Curriculum for Robotic Reinforcement Learning

Manan Tomar, Akhil Sathuluri, Balaraman Ravindran
Indian Institute of Technology Madras, Chennai, Tamil Nadu, 600036, India, +91-9899687343
manan.tomar@gmail.com, akhilsathuluri@gmail.com, ravi@cse.iitm.ac.in

Abstract

Generating a curriculum for guided learning involves subjecting the agent to easier goals first, and then gradually increasing their difficulty. This work takes a similar direction and proposes a dual curriculum scheme for solving robotic manipulation tasks with sparse rewards, called MaMiC. It includes a macro curriculum scheme which divides the task into multiple subtasks followed by a micro curriculum scheme which enables the agent to learn between such discovered subtasks. We show how combining macro and micro curriculum strategies help in overcoming major exploratory constraints considered in robot manipulation tasks without having to engineer any complex rewards and also illustrate the meaning and usage of the individual curricula. The performance of such a scheme is analysed on the Fetch environments.

Introduction

Starting to learn for simpler tasks and then using the acquired knowledge to learn progressively harder tasks is a natural outcome of formulating a curriculum. Recently, curriculum learning has been used to solve complex robotic tasks such as in (Florensa et al. 2017), (Nair et al. 2017). However, these approaches make the assumption that the agent can be reset to any desired state, and also make use of expert state action trajectories (Nair et al. 2017), which are expensive to generate.

The proposed approach, MaMiC, introduces two schemes, macro and micro curriculum, which can be applied either individually or in combination. A micro curriculum essentially generates increasingly complex goals for the agent to achieve. For example, in learning to push a block, initial goals will be generated very near to the block and then slowly shifted to the desired location. However, such a scheme is not sufficient if we need to solve tasks which are more complex, such as ones which require the agent to execute particular sequences of temporally extended actions or sub policies. In order to put an object in a drawer, it is not enough to guide the agent in learning to put the object to the desired location, but also to open the drawer first. It is only when a particular sequence is followed that we refer to the task as completed. A macro curriculum helps in identifying such a sequence and allows the micro scheme to learn in between important subgoals of such a sequence. The policy starts from a sub-goal and proceeds to the next sub-goal, evolving in the process, ultimately reaching the `desired goal`. Two ideas are at the core of this technique, of being able to discover the subgoals and of learning between the recognized sub-goals. Moreover, the Q function $Q(s, a, g)$ and the policy $\pi(s_t, g_t)$ are explicitly parametrized by the goal in order to aid learning in sparse reward settings.

Micro Curriculum

A micro curriculum tries to alleviate the above mentioned assumption of being able to start some trajectories from favourable states. We believe that starting at a particular state should be based on the environment's choice but not the agent's. We propose replacing all or some transition sample goals with `micro goals`, which are artificially generated, using any generative modelling technique. Using an off policy RL algorithm allows us to replace sampled transition goals from the buffer with `micro goals` during learning. The goals are generated such that they are initially close to the achieved states at the end of each trajectory (i.e. the `achieved goal` distribution) and slowly shift to being closer to the actual or `desired goal` distribution of the task in hand. Since this procedure involves learning a mapping between goals and actions, eventually the agent is able to generalize well for the actual goal distribution. We relate this with curriculum learning because the agent initially learns for a goal distribution much simpler to learn i.e. the `achieved goal` distribution and then continues learning for increasingly difficult goals, leveraging the previously learned skills. To train the goal generator, we modify the formulation used by (Held et al. 2017), by incorporating an additional parameter $\alpha \epsilon [0, 1]$ which governs the resemblance of the generated distribution to the `achieved goal` distribution and the actual or `desired goal` distribution. $\alpha = 0$ forces the generator to produce goals similar to the currently achieved states, while $\alpha = 1$ produces goals similar to the actual distribution. The exact objective function is given below.

$$min_D V(D) = \mathbf{E}_{g \sim p_{data}(g)}[(1-\alpha)\ (D(g_{achieved}) - 1)^2 + \alpha\ (D(g_{desired}) - 1)^2]\ +\ \mathbf{E}_{z \sim p_z(z)}[D(G(z))^2]$$

$$min_G \; V(G) \; = \mathbf{E}_{z \sim p_z(z)}[(D(G(z)) \; - 1)^2]$$

,where D denotes the discriminator network, G the generator network, and V the GAN value function.

Strategy for Goal Sampling For replacing goals by sampling new ones, we consider different strategies such as having a mixture of `HER goals` (Andrychowicz et al. 2017) and `micro goals` (referred to as micro-g), only having `micro goals` and having a mixture of `HER goals` and `desired goals` (referred to as micro-sg).

Macro Curriculum

We consider long horizon tasks and assume that few demonstration state trajectories $\tau = s_0, s_1, ...s_T$ are available for the given tasks. Using these demonstration state trajectories, we wish to extract useful `subgoals` of the task. The dense reward $||g_{achieved} - g_{desired}||^2$ per time step for a demonstration is used as the signal for subgoal extraction. The intuition for finding a good sub-goal in a typical manipulation task is to observe that there is a sudden change in the dynamics of the system. For example, if the robot is trying to push a block, it can be easily seen that once the robot explores and starts to interact with the block, the policy will differ as the block interaction dynamics also affect the reward now. For demonstration trajectories, we observe that the gradient ratio of the dense reward always results in consistent spikes near the object's position, proving that it is a good subgoal for learning the tasks such as pushing a block. Additional information is provided in the Appendix (Sec 5). Learning between two such `subgoals` can be performed by following a micro curriculum scheme detailed above. The extracted `subgoals` form a set of states that are achieved by most of the sampled expert trajectories. Note that these `subgoals` are dependent on the start state. Given a policy $\pi(s_t, sg_{t+1})$ that has learnt to achieve a subgoal sg_t allows the agent to achieve the next subgoal sg_{t+1}.

Experiments and Results

Push-hard, Slide-hard, Pick and Place Tasks We consider variants of the pushing, sliding and pick and place tasks for a 7 DOF Fetch robot simulation. These experiments are performed by using micro curriculum for both micro-g and micro-sg sampling strategies.

Receptor-PickAndPlace Task We introduce a new task setting called Receptor-PickAndPlace which comprises of an object placed on a table, a receptor site on the table, and a target located in the air. The agent is required to pick and place the object at the target, which gets activated only if the object passes through the receptor site. Therefore, the agent is not rewarded even if the object is successfully placed at the target, if it does not pass from the receptor site. Such a task is extremely difficult to solve because of sequence of actions involved and a sparse reward available. We show how combining the macro and micro schemes can solve this task, by 1) leveraging demonstration states to extract a subgoal near the receptor and 2) using a powerful micro scheme to realize the sequencing of tasks involved, i.e. first moving the block to the receptor and then to the target. Median success rates for all tasks are shown in the table below.

Task	Micro-sg	Micro-g	HER	MaMiC
Push-hard	100%	92%	1%	-
Slide-hard	42%	31%	1%	-
PickAndPlace	98%	95%	0%	-
Receptor-PickPlace	2%	1%	0%	98%

Results We are able to learn successful policies for all four tasks. For push-hard and slide-hard tasks, the HER baseline does not even learn to reach the object. This can be attributed to a mismatch in the kind of goals provided to the parametrized policy and the ones on which the agent learns off-policy. For Pick and Place, since the goal is always in the air and the object always on the table, a similar mismatch is conceivable. For the Receptor-PickAndPlace task, recognizing the receptor as a subgoal is crucial to learning. There is a significant peak in the dense reward gradient around the receptor location, proving that the subgoal extraction in the macro scheme is able to leverage demonstrations efficiently. This when combined with a micro scheme is able to learn the sequence of going to the receptor first with the block, thus activating the target, followed by placing it over the target. HER and micro scheme applied individually fail to learn this task. We present ablation studies in the Appendix (Sec 7).

Conclusion

We introduce a dual curriculum scheme for robotic manipulation which aids in exploration in tasks with very sparse rewards. We show how the micro scheme is a powerful method for generating goals intelligently and can allow solving hard variants of the pushing, sliding and pick and place tasks without resetting to arbitrary states, starting from favourable states or using expert actions. Moreover, through the Receptor-PickandPlace task, we emphasize on the need for a macro scheme combined with micro when a task involves completing subtasks sequentially.

References

Andrychowicz, M.; Wolski, F.; Ray, A.; Schneider, J.; Fong, R.; Welinder, P.; McGrew, B.; Tobin, J.; Abbeel, O. P.; and Zaremba, W. 2017. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 5048–5058.

Florensa, C.; Held, D.; Wulfmeier, M.; and Abbeel, P. 2017. Reverse curriculum generation for reinforcement learning. *arXiv preprint arXiv:1707.05300*.

Held, D.; Geng, X.; Florensa, C.; and Abbeel, P. 2017. Automatic goal generation for reinforcement learning agents. *arXiv preprint arXiv:1705.06366*.

Nair, A.; McGrew, B.; Andrychowicz, M.; Zaremba, W.; and Abbeel, P. 2017. Overcoming exploration in reinforcement learning with demonstrations. *arXiv preprint arXiv:1709.10089*.