Research Proposal
Learning Complex and Long-Horizon Tasks with Hierarchical Reinforcement Learning

Zhongxuan Li

1 Abstract

One of the most fundamental problems in the intersection of artificial intelligence and robotics is how to enable artificial agents with the ability to automatically composite high-level behaviours from existing motion skills. Interestingly, humans have the innate ability to decompose complex tasks into simple motions in a versatile and agile manner, even in unseen environments. Therefore, to exploit human-like behaviour in robot manipulations is a promising research topic. In this proposal, we review prevalent methods of efficient robot learning within the context of large and unstructured state spaces and sparse extrinsic rewards. Then we discuss our proposed methodology using theoretical framework of hierarchical reinforcement learning together with control-based movement primitives. Finally, we give an overview to expected results and limitations.

2 Introduction

There have been numerous research work on robot learning of individual component skills. Learn From Demonstrations (LfDs) methods such as Dynamic Movement Primitives (DMPs) [1], Probabilistic Movement Primitives (ProMPs) [2] and Task-Parameterized Gaussian Mixture Model [3] aim to learn an optimal robot control policy which is compliant to the distribution of trajectory demonstrations. Based on these approaches, robots can well perform tasks such as grasping and lifting. However, learned motion primitives cannot perfectly execute long-horizon and versatile tasks. Imagine a robot has learned two discrete skills: (1) lifting bricks and (2) carrying bricks. Can it transfer its skill to a complete high-level task of moving heavy loads from one point to another, or even recover from failures? And can it generalize its ability to different shapes of bricks? To solve the aforementioned questions of both low-level motion planning and high-level task planning, not only do we need to decide the continuous dynamic movement parameters, but we also need to determine the scheme to composite different motion skills into long-horizon behaviours. It is worth noticing that manipulation tasks often have a inherent modular structure that may be used to improve performance across a family of tasks. Therefore, previous research have attempted to break down a manipulation task into component motor skills. Decomposing tasks in this approach has several advantages, based upon which Hierarchical Reinforcement Learning (HRL) is introduced to allow agents to demonstrate behaviours of increasing complexity through continuously building on its existing skill library. Because each skill has a shorter-horizon, HRL reduces the complex robot learning problem to substantially easier learning problem and aiding exploration. Furthermore, by decomposing a high-level robot behaviour into low level motor

skills, we have the access to abstractions over the input that robot uses to decide which skill to execute. This mechanism also makes learning for new tasks more effective by enabling generalized policies that work across tasks. Moreover, the resulting abstract representations may be much accessible for a non-expert user to understand and edit esoteric robot behaviours. Hence applying HRL on robot learning can enhance safety and trust in human-machine systems, which is valuable for developing Accountable, Responsible and Transparent AI.

3 Literature Review

The crux of hierarchical structure in robot manipulation learning has two major problems: (1) Identifying the motion primitives from which a solution are likely to be assembled. This can be achieved either from explicit demonstrations, or from behaviors generated by the robot while solving tasks. (2) Learning a high-level options framework that is used to decided which skill to execute next, based on current state-action parameters.

Learning From Demonstrations. The most prevalent approach for learning motion skills from demonstrations is Dynamic Movement Primitives (DMPs) which encode motion policy in mathematical formulation, based on second-order dynamic systems[4]. Probabilistic Movement Primitives (ProMPs) [2] extend DMPs to a probability distributions framework that enables adaptation to a new task or a new situation with altered desired trajectories. There are various works on applying movement primitives to robot manipulation tasks such as stir-fry [5], folding assembly [6]. and impedance control [7]. Motor primitives could be combined with reinforcement learning for better parameter tuning, providing several rollouts to find a better policy update. An exemplary work is Policy Learning by Weighting Exploration with the Returns (PoWER) [8] that combined reinforcement learning with motor primitives for parameter tuning based on expectation maximization. All these approaches need a complete task description or strong human supervision, thus limit the scalability of these methods in realistic tasks.

Discovering Skills While Solving Tasks. An alternative approach is to discover component skills during the process of learning to solve manipulation tasks. The key is to identify the skills from a group of continuous longer trajectories, which is an under specified challenge. This can be triggered by either a salient signal or to reach another skill's initiation conditions. After a family of motor skill segments is collected, a measure of skill similarity can be applyed to group up demonstration trajectories. The most intuitive method it to measure the skills similarity by fitting the data to a parameterized policy domain and measuring distance metric in that parameter space [9]. Apart from adopting a parametric model, a latent space representation can be discovered to effectively encodes skills. Other recent methods rely on observations only, even from videos of decomposing multi-step tasks into composable primitives [10].

Learning Decision-Making Abstractions. After acquiring a skill library, a high-level policy network is learned to decided which skill to execute next, based on current state-action parameters. The Option-Critic architecture [11] devised a gradient based framework to generate options during learning the policy. Option, which is a tuple composed of an initiation set, a policy, and a termination condition, are automatically generated by the agents itself, leaving the number of options the only hyper-parameter to be tuned. However, option-based framework are often defined in the latent policy search space, thus does not fully exploit the benefits of expert trajectories, making it difficult to identify useful sub-goals for complex tasks. Instead, the sub-goal based hierarchical framework [12] enables complex real-world tasks to be decomposed into sub-goals in some natural temporal order. For example, in [13] and [14], the whole framework is decomposed into a state-independent task schema (sequence of skills) and a state-dependent policy. However, these kind of methods need careful reward definition over sub-task termination

conditions, making them hard to generalize.
To conclude this section, movement primitives can often be quickly acquired from demonstrations using behavioural cloning or apprenticeship learning, while mastery of manipulation skills requires additional training and can be achieved using reinforcement learning. In the future research, we intend to apply the HRL framework in solving some robotic manipulation tasks, all within the context of real-world robots.

4 Proposed Methodology

We extend this idea of HRL further in this proposal, giving an overview of a possible approach to solving long-horizon, sparse reward tasks given a set of parameterized skills by learning a policy that chooses which skill to execute and what arguments to use when invoking it. As robots and humans interact intimately, the essence of interpretability of robot decision-making processes has higher eminence. Mutual understanding of underlying representations of both the environment and task is a prerequisite for successful human-robot collaboration. Therefore, in this proposal, we tend to decompose the entire complex task into explainable sub-goals and try to reach them in the best sequential order. We specifically focus on utilizing both learning-based methods (reinforcement learning) and model-based methods (movement primitive). We first train several predefined movement primitives by imitation learning. The training trajectories can be acquired through Virtual Reality motion capture devices or motion retargeting from videos. Because each motion primitive has only limited scope in terms of duration and the tasks it can achieve. Therefore, after collecting a library of skills, we train two policy networks for a robot arm via reinforcement learning: A low-level policy networks determine skill parameters that automatically adapt to current continuous state-space arguments and a prespecified sub-goal. In addition, a high-level policy network is trained to decompose a long-horizon task into short sub-goals, which serve as input to the low-level policy network. By utilising this compositional structure of planning to generate diverse behaviours, we are able to reduce the difficulty of training a diverse goal proposal model. Note that, our proposed method differs from fully model-free methods in [15] by the fact that both high-level and low-level controllers in our framework are defined by parameters that are meaningful in physics. Instead of training the low-level policy using model-free methods such as Deep Deterministic Policy Gradient (DDPG), we use dynamic movement primitives or other model-based methods, enhancing the explainability of our low-level policies.
Problem Setup. A manipulation task can be approximate to a sequential decision making problem and modelled as a discrete-time infinite-horizon Markov Decision Process (MDP), $M = (S, A, T, R, \gamma, \rho_0)$, where S is the state space, A is the action space, $T(\cdot|s, a)$ is the state transition distribution, $R(s, a, s')$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and $\rho_0(\cdot)$ is the initial state distribution. Formally, we intend to learn a (1) low-level policy network $\pi_\theta(s, s_g)$, that outputs trajectories $\tau_{1:H}$ to try and reach a goal observation s_g that is H timesteps away, and (2) a high-level goal proposal network, $p_\theta(s_g|s)$, that samples goals for the low-level network to reach. Both models are trained on sequence of demonstrated trajectories. The skill sets is define as follows. $\tau^2\ddot{y} = \alpha y(\beta y(g - y) - \tau\dot{y}) + f$, $f(x, g) = \sum_{i=1}^{N} \varphi_i\omega_i / \sum_{i=1}^{N} \varphi_i x(g - y_0)$, where y is the joint position vector and g is the target pose, $f(x)$ is defined as a linear combination of N nonlinear Radial Basis Functions (RBFs). At each sub-goal transition state, $g = s_g$. The best policy (i.e. MP parameters) can be found by stochastic optimization. Note that, the goal parameter g of a primitive is also the start parameter y_0 of the next primitive in the sequence, and its execution can affect the cost of the subsequent primitives.
Equipments needed. To evaluate our methods in both simulation environment and physical world. We need collaborative robots designed specifically for research purposes. For example,

Franka Emika Panda [16], a 7-dof arm which is good at torque control and human-robot interaction. Some other options may include the UR5 collaborative robot arm [17]. These arms are designed to assist humans in various tasks, such as manipulation. They are installed with force sensors, allowing it to automatically freeze if something unexpected enters the workspace. In addition, we may need to exam our methods on a mobile manipulator, available options include the Fetch Robot [18] and Hello Robot Stretch[19]. All these robots are being widely used throughout the robotics research community, making it easy to compare our results with the baseline methods. As for simulation environment, there exist several differentiable physics simulators (e.g. Mujuco, iGibson) embedded with the OpenAI Gym framework for better training the model.

5 Limitations

The state space parameter of the policy network is composed of high-dimensional observations and sub-goals. In high-dimensional observations domains such as images, this may require explicitly optimizing over image pixels, which is undesirable. Hence it is crucial to find latent expressions of the image signals, similar in [20]. For generality, the lower-level controllers should be supervised with goals that are learned and proposed automatically by the higher-level controllers [21], without any human prior knowledge. Hence we shall look into the mechanism of segmenting the overall goal into meaningful subtasks. In addition, training the model in simulators requires millions of frames of interaction until convergence and the learned policy is hard to generalize to drastically different settings. We need to pay attention to how to narrow the sim2real gap.

6 Expected Results

We plan to evaluate the proposed method on variations of manipulation tasks both in simulation environment and on physical robots. The conducted experiments could be based on tasks introduced by [22] and on variations of other problems. The simulation utilize the Mujoco based OpenAI Gym environment. Two environments FetchPush-v1 and FetchPickAndPlace-v1 could be selected as baseline tasks, where the robot need to reach and pick different types of objects in long-horizon movements, which has sparse and binary reward functions. The generalization to new target domains could be tested with different object starting locations, and varies the object's coefficient of friction. In addition, we evaluate the same experimental setup in a physical environment, to better exam the ability of transferring robot skills acquired in simulation to the real robotic system. We can evaluate the results in multiple aspects, including the rate of convergence and the success rate of execution in both seen and unseen scenarios.

7 Conclusion

Hierarchical reinforcement learning in robot manipulation have shown a promising future in many manufacturing, assembly lines, and household processes. Our work demonstrates the potential to take this intelligence one step further, that combing the control-based movement primitives with hierarchical reinforcement learning to generate fully automated policies for complex and long-horizon tasks. We expect to obtain results that excel the baseline and we intend to investigate the idea further in future studies.

References

[1] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: Learning attractor models for motor behaviors," *Neural Computation*, vol. 25, no. 2, pp. 328–373, 2013. 2

[2] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann, "Probabilistic movement primitives," in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, 2013. 2, 3

[3] S. Calinon, Z. Li, T. Alizadeh, N. G. Tsagarakis, and D. G. Caldwell, "Statistical dynamical systems for skills acquisition in humanoids," in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pp. 323–329, 2012. 2

[4] M. Saveriano, F. J. Abu-Dakka, A. Kramberger, and L. Peternel, "Dynamic movement primitives in robotics: A tutorial survey," *CoRR*, vol. abs/2102.03861, 2021. 3

[5] J. Liu, Y. Chen, Z. Dong, S. Wang, S. Calinon, M. Li, and F. Chen, "Robot cooking with stir-fry: Bimanual non-prehensile manipulation of semi-fluid objects," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5159–5166, 2022. 3

[6] D. Almeida, F. E. Viña, and Y. Karayiannidis, "Bimanual folding assembly: Switched control and contact point estimation," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pp. 210–216, 2016. 3

[7] A. Batinica, B. Nemec, A. Ude, M. Raković, and A. Gams, "Compliant movement primitives in a bimanual setting," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pp. 365–371, 2017. 3

[8] J. Kober and J. Peters, "Policy search for motor primitives in robotics," in *NIPS*, pp. 849–856, 2008. 3

[9] C. Daniel, H. van Hoof, J. Peters, and G. Neumann, "Probabilistic inference for determining options in reinforcement learning," *Machine Learning*, vol. 104, 09 2016. 3

[10] W. Goo and S. Niekum, "Learning multi-step robotic tasks from observation," *CoRR*, vol. abs/1806.11244, 2018. 3

[11] R. S. Sutton, D. Precup, and S. Singh, "Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning," *Artificial Intelligence*, vol. 112, no. 1, pp. 181–211, 1999. 3

[12] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. B. Tenenbaum, "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation," *CoRR*, vol. abs/1604.06057, 2016. 3

[13] S. Nasiriany, H. Liu, and Y. Zhu, "Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 7477–7484, 2022. 3

[14] R. Chitnis, S. Tulsiani, S. Gupta, and A. Gupta, "Efficient bimanual manipulation using learned task schemas," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1149–1155, 2020. 3

[15] B. Beyret, A. Shafti, and A. A. Faisal, "Dot-to-dot: Achieving structured robotic manipulation through hierarchical reinforcement learning," *CoRR*, vol. abs/1904.06703, 2019. 4

[16] F. Emika, "Franka research cobot (https://www.franka.de/)." 4

[17] U. Robot, "Universal robot (https://www.universal-robots.com/products/ur5-robot/)." 4

[18] F. M. Manipulator, "Fetch mobile manipulator (https://www.universal-robots.com/products/ur5-robot/)." 4

[19] H. Robot, "Hello robot stretch re1 (https://hello-robot.com/product)." 4

[20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014. 5

[21] O. Nachum, S. Gu, H. Lee, and S. Levine, "Data-efficient hierarchical reinforcement learning," *CoRR*, vol. abs/1805.08296, 2018. 5

[22] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, V. Kumar, and W. Zaremba, "Multi-goal reinforcement learning: Challenging robotics environments and request for research," *CoRR*, vol. abs/1802.09464, 2018. 6