ETH zürich

CRL Computational Robotics Lab

Learning Agile Robotic Behaviors via Zero-Shot Reinforcement Learning

Motivation

Reinforcement learning (RL) provides a framework to obtain optimal or near-optimal policies from sub-optimal data given a reward function. In recent years, the use of RL to control quadruped robots has achieved significant success, enabling robust locomotion across highly diverse terrains [1, 2, 3] and the acquisition of diverse behaviors [4, 5]. However, it is infeasible to enumerate all possible reward functions which may be of interest to solve in the future, and hence most RL approaches rely on fixed rewards for training, limiting the generalizability of the learnt policies to new tasks.

Figure 1: Potential behavior obtained in a zero-shot manner

Zero-shot RL [6, 7, 8] seeks to address this limitation by learning optimal policies for all possible reward functions. In this way, an agent may, with a minimal amount of extra computation, infer an optimal policy for *any reward function given at test time*.

One of the most classic instantiations of zero-shot RL is goal-conditioned methods [9, 10], which train goal-conditioned policies to reach any goal state from any other state. However, these methods are restricted to goal-reaching tasks only.

Recent work has introduced forward-backward (FB) representations [8], which aims to factorize the occupancy distribution of the policies into a forward representation (F) of the current state and backward representation (B) of a target state. One of the limitations of such methods is the reliance on datasets with good coverage. This problem has been tackled by using exploration policies trained with an intrinsic exploration reward [11, 12, 13, 14]. More recently, Urpí et al. [15] proposed an exploration strategy for efficiently learning the representations online by minimizing the epistemic uncertainty on the learned representations. Additionally, Tirinzoni et al. [16] biased the exploration towards relevant states by regularizing unsupervised RL towards imitating trajectories from an unlabeled behavior dataset.

Nonetheless, none of these techniques have yet been leveraged to real robotic systems. This project aims to implement FB method and evaluate its performance on the quadruped robot, Unitree Go2 (see fig. 1). Excitingly, at test time, the learned model can be prompted to solve entirely new tasks—such as walking with a specific gait, tracking a desired motion, reaching a target pose, or even performing a backflip without requiring any additional learning or fine-tuning. This highlights the promise of zero-shot RL in enabling highly flexible and adaptable robotic behavior, making it an exciting and impactful area for further exploration.

Goal

The objective of this project is to implement the FB method on a real quadruped robot, benchmark it against other zero-shot RL algorithms, and analyze the advantages and limitations of FB in real-world robotic scenarios.

Requirements

The student brings along the following attributes:

1. Proficiency in Python and machine learning (familiar with PyTorch).
2. Good knowledge of reinforcement learning.
3. Experience with hardware is preferred but not essential.
4. Motivation for the project

Interested?

We look forward to working with motivated students who are passionate about reinforcement learning and robotics. Reach out to Núria Armengol (nuria.armengolurpi@inf.ethz.ch) and Jin Cheng (jin.cheng@inf.ethz.ch) with your CV and transcripts.

References

[1] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):eabc5986, 2020.

[2] Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62): eabk2822, 2022.

[3] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning*, pages 91–100. PMLR, 2022.

[4] Tom Zahavy, Yannick Schroecker, Feryal Behbahani, Kate Baumli, Sebastian Flennerhag, Shaobo Hou, and Satinder Singh. Discovering policies with domino: Diversity optimization maintaining near optimality. In *The Eleventh International Conference on Learning Representations*.

[5] Jin Cheng, Marin Vlastelica, Pavel Kolev, Chenhao Li, and Georg Martius. Learning diverse skills for local navigation under multi-constraint optimality. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5083–5089. IEEE, 2024.

[6] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993. doi: 10.1162/neco.1993.5.4.613.

[7] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[8] Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.

[9] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

[10] Vitchyr H Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning. *arXiv preprint arXiv:1903.03698*, 2019.

[11] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[12] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

[13] Hao Liu and Pieter Abbeel. Aps: Active pretraining with successor features. In *International Conference on Machine Learning*, pages 6736–6747. PMLR, 2021.

[14] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pages 2778–2787. PMLR, 2017.

[15] Núria Armengol Urpí, Marin Vlastelica, Georg Martius, and Stelian Coros. Epistemically-guided forward-backward exploration. In *Reinforcement Learning Conference*.

[16] Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirotta. Zero-shot whole-body humanoid control via behavioral foundation models. *arXiv preprint arXiv:2504.11054*, 2025.