ES-sim2real optimization

Giacomo Spigler *g.spigler@tilburguniversity.edu*; Cesare Maria Dalbagno *c.m.dalbagno@tilburguniversity.edu*
AI for Robotics Lab (AIR-Lab), Department of Cognitive Science and Artificial Intelligence
Tilburg University, Tilburg, The Netherlands

Introduction

Reinforcement learning (RL) has emerged as a promising approach for enabling robots to acquire complex skills and adjust to new surroundings. Nevertheless, the process of training RL controllers directly on real robots presents significant challenges due to the extensive amount of experience RL algorithms demand, making it more practical to train them in physics simulators (e.g., Rajeswaran et al., 2017).

However, behaviors that perform well in simulations often struggle to translate effectively to the real world. This is largely due to the 'sim-to-real gap', that is, an inevitable mismatch between even the most accurate simulations and actual robots leading to a series of complications (Zhao et al., 2020). A modern approach to improve the effectiveness of transferring skills from simulations to the real world (sim2real transfer) utilizes domain randomization (Weng, 2019). This method trains agents across a wide array of simulations featuring varied and randomized physics parameters. The aim is to broaden the range of environments in which the agent can effectively apply its learned skills, expecting that the real world resembles one of these simulations closely. It also helps prevent the agent from overfitting to a particular set of parameters.

To approach the sim2real challenge, we propose to adopt a meta-learning framework explicitly to train simulated agents to perform few-shot adaptation to novel physical environments, with a framework close to that of (Song et al., 2020). We model a distribution over possible physics using domain randomization, where each configuration is essentially treated as a distinct possible task. Sim2real is then performed as few-shot adaptation, optimized for using state-of-the-art meta learning algorithms (MAML (Finn et al., 2017), and in particular its derivative-free variant ES-MAML (Song, Gao, et al., 2020), based on evolution strategies), with gradient-based inner-loop optimization performed using Proximal Policy Optimization (PPO; Schulman et al., 2017).

The novel contributions of our work are as follows: (1) we adapt previous work (Song et al., 2020) to use PPO as gradient-based inner-loop adaptation operator; (2) we explore visualizations of the `sim-to-real gap' across a variety of baselines, following the analysis from (Rajeswaran et al., 2016); (3) we perform an extensive sim2sim empirical evaluation of the method [work in progress], comparing it against a typical domain randomization baseline that learns a meta-learning policy implicitly through an LSTM; (4) [work in progress] we further extend the method to enable meta-learning a reward function that relies on the same observations as the policy network, in a way similar to Evolved Policy Gradients (EPG; Houthooft et al., 2018), to enable adaptation on real robots without requiring hand-crafting reward functions and complicated setups in the real world.

Methods

[The project is currently work-in-progress; most of the simulations are being setup, and should be completed by May]

Analysis of the performance of various methods in controlled sim2sim transfer settings was performed on the `Hopper-v4' Mujoco environment (Erez et al., 2012; Todorov et al., 2012; Towers et al., 2023). We trained a fully connected neural network using PPO from stable-baselines3 (Raffin et al., 2021). The task the one-legged robot is trying to achieve is to move as quickly as possible in one direction without falling. We have tested performances in three distinct learning conditions. In the first one we used the default values for the *torso mass* and *ground friction* parameters in the environment (mass=6, friction=2); in the second we realized a simple domain randomization by sampling the parameters from two uniform distributions within [1, 9] (mass) and [1.5, 2.5] (friction) and lastly we implemented ES-MAML (500 meta iterations, with inner-loop adaptation through 3 PPO updates each with 4096 environment steps) in the same randomization setup. All three models were later tested in an environment with parameters outside of the training distribution (mass range = [0.5, 11.5]; friction range = [0.1, 2.9]) to emulate a distribution shift.

Results and conclusions

Preliminary results are reported in Figure 1, averaged over 5 random seeds. While traditional domain randomization significantly improves generalization performance compared to the fixed parameters baseline, it is found to remain limited to the specific range of physics parameters it was trained on, with only limited generalization outside the distribution. Our proposed method instead shows promising out-of-distribution generalization.

Even though the full evaluation is still a work in progress, based on the experiments conducted until now, we are confident that the further implementation of a learnt reward function will lead to even better outcomes, providing a significant contribution to the field.

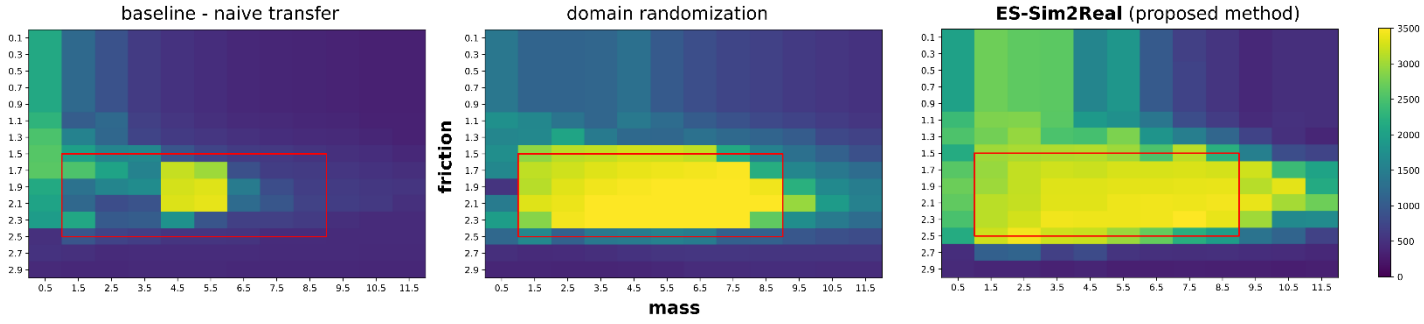

Figure 1:

Baseline-naive-transfer agent: trained in a fixed physics simulator
Domain randomization: trained over a collection of simulated, the range of randomization parameters is shown as a red box in the plots.
Proposed method (ES-Sim2Real): agents are trained using meta-reinforcement-learning

References

Erez, T., Tassa, Y., & Todorov, E. (2012). *Infinite-Horizon Model Predictive Control for Periodic Tasks with Contacts*. https://doi.org/10.7551/mitpress/9481.003.0015

Finn, C., Abbeel, P., & Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning (pp. 1126-1135). PMLR.

Houthooft, R., Chen, Y., Isola, P., Stadie, B., Wolski, F., Jonathan Ho, O., & Abbeel, P. (2018). Evolved policy gradients. Advances in Neural Information Processing Systems, 31.

Pinto, L., Andrychowicz, M., Welinder, P., Zaremba, W., & Abbeel, P. (2017). Asymmetric actor critic for image-based robot learning. arXiv preprint arXiv:1710.06542.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*, *22*(268), 1–8. http://jmlr.org/papers/v22/20-1364.html

Rajeswaran, A., Ghotra, S., Ravindran, B., & Levine, S. (2016). Epopt: Learning robust neural network policies using model ensembles. arXiv preprint arXiv:1610.01283.

Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., & Levine, S. (2017). Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. arXiv preprint arXiv:1709.10087.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal Policy Optimization Algorithms* (arXiv:1707.06347). arXiv. https://doi.org/10.48550/arXiv.1707.06347

Song, X., Gao, W., Yang, Y., Choromanski, K., Pacchiano, A., & Tang, Y. (2020). *ES-MAML: Simple Hessian-Free Meta Learning* (arXiv:1910.01215). arXiv. https://doi.org/10.48550/arXiv.1910.01215

Song, X., Yang, Y., Choromanski, K., Caluwaerts, K., Gao, W., Finn, C., & Tan, J. (2020). Rapidly Adaptable Legged Robots via Evolutionary Meta-Learning. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3769–3776. https://doi.org/10.1109/IROS45743.2020.9341571

Todorov, E., Erez, T., & Tassa, Y. (2012). MuJoCo: A physics engine for model-based control. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. https://doi.org/10.1109/IROS.2012.6386109

Towers, M., Terry, J. K., Kwiatkowski, A., Balis, J. U., Cola, G., Deleu, T., Goulão, M., Kallinteris, A., KG, A., Krimmel, M., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Tan, A. J. S., & Younis, O. G. (2023). *Gymnasium* (v0.29.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.8269265

Weng, L. (2019, May 5). *Domain Randomization for Sim2Real Transfer*. https://lilianweng.github.io/posts/2019-05-05-domain-randomization/

Zhao, W., Queralta, J. P., & Westerlund, T. (2020). Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: A Survey. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 737–744. https://doi.org/10.1109/SSCI47803.2020.9308468