ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Master Thesis

Identifying surprise with conditional computation

Reinforcement Learning (RL) has gained a lot of attention in recent years. In RL an agent interacts with an environment through actions and gets rewards when certain conditions are met. The goal of the agent is to learn the sequences of actions that maximize the reward it gets. This approach has been successfully used to learn all kinds of different tasks, from learning to play Atari games [1] to complex robotic control [4].

Current techniques very often rely on random exploration. That is, the agent performs random actions in the environment to explore which policies (mapping of environment states to actions) return the highest reward. While this alone tends to be enough in many scenarios, it quickly becomes unfeasible as environments become more complex, as the probability of finding the right set of actions drops exponentially.

To counteract this, smarter ways of doing exploration have been proposed in the literature. These techniques try to find states or transitions of states that are 'interesting' and guide the agent to explore those. The way this is done is either by counting the number of times the current state has been visited [2] or trying to predict the next state [3]. These approaches have been shown to be extremely successful in deterministic environments . However, these approaches suffer from the noisy TV problem. That is, if the agent has access to a set of states that output random noise each time (i.e. a noisy TV), they will find these states extremely interesting as they cannot be predicted and are completely new each time. Thus, the agent gets stuck 'watching' the noisy TV and never actually solves the original task.

In this thesis you will explore how to find surprising transitions without getting stuck on the noisy TV. In particular we will look at the use of conditional computation for this. This allows us to focus on how the transition should be processed, rather on what the transition will be. We have built a preliminary model and shown in some toy problems that we can get rid of the noisy TV problem. Your goals in this thesis will be:

- Familiarize yourself with state-of-the-art exploration methods.
- Come up with different simple environments where standard exploration approaches fail.
- Experiment with different conditional computation techniques like discrete gating and hyper-networks on toy problems. This will help better understand the model and thus, improve it and make it more reliable.
- Extend the approach to more complex and realistic environments, like Atari or 3D mazes.

More information and grading scheme can be found on:
`https://www.cadmo.ethz.ch/education/thesis/guidelines.html`

Prerequisites: Knowledge of Tensorflow or Pytorch

Supervisor: Asier Mujika, CAB J 21.2, `asierm@inf.ethz.ch`

Supervising Professor: Prof. Dr. Angelika Steger, CAB G 37.2, `steger@inf.ethz.ch`

References

[1] A. P. Badia, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, Z. D. Guo, and C. Blundell. Agent57: Outperforming the atari human benchmark. In *International Conference on Machine Learning*, pages 507–517. PMLR, 2020.

[2] Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

[3] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pages 2778–2787. PMLR, 2017.

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.