Self-Imitation Learning[*]

Junhyuk Oh, Yijie Guo et al.

2018

1 What

- Self-Imitation Learning (SIL) algorithm which learns to imitate agent's good past decisions.
- Theoretical justification of SIL (SIL's objective is derived from a lower bound of the optimal soft Q-function) within an entropy-regularised RL framework.
- Showing that A2C+SIL does a good job on hard exploration games as well as improves overall performance on 49 Atari games.
- Showcase of SIL+PPO.

2 Why

Exploration/exploitation dilemma is old as the world, but it's still an open question. What if we exploit more from successful past experiences of the agent itself?

3 How

SIL flow:

- Play an episode till the end and store the experiences
- Recalculate the return and store it in the experience replay instead of reward
- Do an A2C update on the on-policy samples

[*] Notes by Vitaly Kurin `https://yobibyte.github.io/`. Thanks to most of the authors for the paper sources (I copypaste long formulas and figures sometimes). And thanks to you all for feedback. Stay tuned!

- Do several minibatch updates on the prioritised samples from memory replay with the SIL objective

SIL objective encourages an agent to imitate when the return is higher than expected (than $V_\theta(s)$):

$$\mathcal{L}^{sil} = \mathbb{E}_{s,a,R \in \mathcal{D}} \left[- \log \pi_\theta(a|s)(R - V_\theta(s))_+ + \frac{\beta}{2} ||(R - V_\theta(s))_+||^2 \right], \qquad (1)$$

where $(\cdot) = \max(\cdot, 0)$. In addition to this, we want to sample interesting trajectories which have higher clipped advantages. The authors use Prioritized Experience Replay [Schaul et al., 2015] with $(R - V_\theta(s))_+$ as priorities.

As the authors claim, SIL can be used with any actor-critic method. They decided to go with A2C [Mnih et al., 2016].

The paper provides the justifications of the following claim: SIL objective $\mathcal{L}^{sil}$ can be viewed as an implementation of lower-bound-soft-Q-learning under the entropy-regularized RL framework [Ziebart et al., 2008, Haarnoja et al., 2017]. Without any details, any policy is a lower bound of an optimal policy (optimal soft-Q-value in this case). Lower-bound soft-Q learning objective encourages us to update only on those experience which has the Q lower than the return of a soft-Q policy:

$$\mathcal{L}^{lb} = \mathbb{E}_{s,a,R \in \mu} [\frac{1}{2} ||R - Q_\theta(s,a))_+||^2], \qquad (2)$$

where $R_t = r_t + \sum_{k=t+1}^{\infty} \gamma^{k-t}(r_k + \alpha \mathcal{H}_k^\mu)$.

4 Evaluation

I really like that at the beginning of the evaluation, the authors pose the questions which motivate the furthrer experiments:

- Is SIL useful for exploration
- Is SIL complementary to count-based exploration methods?
- Does SIL improve the overall performance across a variety of tasks?
- When does SIL help and when does not?
- Can other off-policy actor-critic methods also exploit good experiences?
- Is SIL useful for continuous control and comparable with other algorithms such as PPO [Schulman et al., 2017].

The author tests their approach in three domains: Key-Door-Treasure, Atari 2600 and MuJoCo.

4.1 Key-Door-Treasure

The first one is a grid-world where the agent should pick up the key, open the door with it and collect the treasure. The Apple-Key-Door-Treasure modification makes the task harder by adding small rewards for collecting apples near the agent. The authors implement [Strehl and Littman, 2005], a count-based exploration method to compare it with SIL. Though, there is not much difference between using both at the same time for the first version of the environment. However, for the environment with apples, it makes a huge difference.

That's an interesting experiment, it looks like my initial understanding was correct: you do not only need to maintain the balance between exploration and exploitation, but you also need to exploit effectively so that it leads to learning useful behaviour. And this combination of the count-based exploration method and SIL works well here.

4.2 Atari 2600

Atari evaluation which I'm the fan of [1] is also interesting here. First, the authors show, that SIL works better than count-based exploration actor-critic methods (A3C-CTC [Bellemare et al., 2016], Reactor-PixelCNN [Ostrovski et al., 2017] and TRPO-SimHash [Tang et al., 2017]) on 6 out of 7 hard exploration games (with one being very close to the baseline): Montezuma's Revenge, Freeway, Hero, Private Eye, Gravitar and FrostBite. It fails to do anything for Venture. According to the authors, it's due to the fact that random exploration doesn't lead us anywhere. We need something more sophisticated to overcome these problems.

The comparison is curious since SIL does not have any sophisticated exploration strategy (apart from entropy regularization in A2C part). It would be also good to see just A2C/A3C in the comparison table for hard-exploration games, since A3C in [Mnih et al., 2016] does really well and achieves results very similar to A2C+SIL. The results on Freeway are also interesting. A3C in [Mnih et al., 2016] can't do anything. At the same time, DQN and DDQN do quite a good job there. Why does it happen?

This is a good example of the fact, that Atari 2600 games are all different and comparing mean/median score is weird. Can we get more insights about our algorithms/exploration strategies from discussing each game case separately (or some subset of them).

Another interesting exploration is that sometimes the method learns faster at the beginning, but gets stuck in a suboptimal policy. So, excessive exploitation hurts. The authors write that reducing the number of SIL updates per iteration helps.

Finally, the paper compares ACER [Wang et al., 2016]+SIL sampling priority and A2C-SIL and ACER+SIL does not do really well. The authors conjecture that ACER objective has an importance sampling term and thus hinders the algorithm from learning good behaviour if the current policy deviates too

[1] `https://yobibyte.github.io/atari-eval.html`

much from old decisions. That's very interesting. Does it mean that caring too much about being on-policy hurts? Or it's rather caring too much about being on-policy when doing an off-policy update.

4.3 MuJoCo

The control part is quite brief, but there is a very interesting experiment here as well. The authors *sparsify* the reward by returning a cumulative reward after every 20 steps. The performance of PPO-SIL vs PPO is not that different, however, the gap in the second experiment with sparser reward is larger. Again, it looks very beneficial when an agent can exploit rare experiences as much as possible [2].

5 Comments

- Interestingly, people usually talking about exploration strategies, not exploitation. Here it's the other way around.
- As far as I understand the motivation, the point is that when we exploit, we do not exploit good enough to repeat past successes. This paper focuses on this repeating 'good enough'.
- The algorithm description is a bit weird. First, it's clear, that the episode should be played till the end to calculate the return afterwards. But then the authors sum the discounted return till the infinity. I'm not sure, but can it break something in theoretical justifications? Or it's all right and we can assume that the finite horizon also work? What should we do if there is no 'terminal' event in the environment? Same horizon case?
- The authors mention *deep exploration* in the conclusion. What's that?
- I'm always very confused about some people being cautious about on-policyness and some people who don't care or do care but just don't mention it. Here, we have two algorithms, one off-policy (A2C), one off-policy (Q-learning). We collect data, do an on-policy update, do off-policy updates. Why doesn't it screw up our policy? We update the same θ in the off-policy update. Sounds very dangerous. But it works! Why?
- The previous point magic might work because of the fact that SIL objective is a lower-bound-soft-Q under the entropy regularised RL framework. Or is it just because we don't use an on-policy update on the off-policy data? And there is nothing wrong with combining all the possible algorithms as soon as we do not do an on-policy update with outdated data.

[2] "Look, if you had one shot, one opportunity to seize everything you ever wanted. One moment. Would you capture it or just let it slip?" `Beat goes on...`

- On the one hand, SIL leads to achieving more in the environments with sparse reward. But what about environments with a lot of distractive rewards? The authors ask this question in the introduction, but I haven't seen an explicit answer to this question later in the paper. Looks like the experiment with apples gives us the answer for that.

References

[Bellemare et al., 2016] Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479.

[Haarnoja et al., 2017] Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*.

[Mnih et al., 2016] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937.

[Ostrovski et al., 2017] Ostrovski, G., Bellemare, M. G., Oord, A. v. d., and Munos, R. (2017). Count-based exploration with neural density models. *arXiv preprint arXiv:1703.01310*.

[Schaul et al., 2015] Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

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

[Strehl and Littman, 2005] Strehl, A. L. and Littman, M. L. (2005). A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*, pages 856–863. ACM.

[Tang et al., 2017] Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, O. X., Duan, Y., Schulman, J., DeTurck, F., and Abbeel, P. (2017). # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2750–2759.

[Wang et al., 2016] Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. (2016). Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*.

[Ziebart et al., 2008] Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA.