Note

Monday, August 10, 2020 9:44 PM

1. I was supposing that by regarding the reachability with more future states from s_t , in the calculation of curiosity bonus can achive better results. (like computing curiosity based on more Reachability(s_t, s_{t+1}), Reachability(s_t, s_{t+2}), Reachability(s_t, s_{t+3}) … .pairs can achieve better results than just based on Reachability(s_t, s_{t+1}). But it turns out to be calculating curiosity based on just Reachability(s_t, s_{t+1}) achieves a better result in the experiments I have tried (Breakout and MountainCar). For other environments like Kangaroo, Seaquest all failed with different combinations of hyperparameters and reachability distances.

2. The most popular curiosity approach is the Random Network Distillation, which models the novelty at s_t based on the difference of $f_1(s_t)$ and $f_2(s_t)$ and we train f_1 to getting close to f_2.

   The advantage of this approach is that it calculates the curiosity of s_t just based on the novelty of s_t instead of the novelty of trainsition pair (s_{t-1}, s_t). The novelty of novelty of trainsition pair (s_{t-1}, s_t) only shows that s_t is novel to s_{t-1}, but doesn't necessary reflect the novelty of s_t. And state-transition-based method can easily be fooled by noisy TV problems, and the Episodic Curiosity Through Reachability aims to solve the nosiy TV problem but it comes with super high computational cost.

3. I have seen others to use VAE, GAN, Attention and Successor Representation to generate curiosity.

4. "Never Give Up: Learning Directed Exploration Strategies" achieves some good results by multiplying the episodic curiosity (similar to the approach in Episodic Curiosity Through Reachability) with the life long curiosity (similar to the approach in Random Network Distillation) with distribution DQN.

5. Problems and Concerns of DQN from My View
   a. The learning of curiosity: it is very slow to help the previous states learn the curiosity reward obtained in the future states in DQN and the learning will be further delayed due to the update of the target network if we use double DQN. And the curiosity reward at state s_t will be continuously decayed as s_t is reached more and more times, the early states may never learn the novelty at the future states that are furthur away from them. However, it may still be important for the early states to learn the novelty of the future states.

   b. DQN is a value-based algorithm, and the curiosity reward is not constant as the extrinsic reward, in my opinion, curiosity reward will contribute to the error in the estimation of values. (But it doesn't seem to cause big problem)

6. My Ideas:

To fix 5.a, can we split the $Q^{combined}(s,a)$ into $Q(s,a)$ and C(s, a) and propagate the novelty of future states with a faster speed for C(s, a)?

$Q(s,a)$ is the state action value of extrinsic reward and $Q(s_t,a_t)$ = R + $argmax_{a_{t+1}}(\gamma Q(s_{t+1},a_{t+1}))$.

And, C(s, a) is the state action value of intrinsic reward and C(s_t,a_t) = Novelty(s_{t+1}) +$argmax_{a_{t+1}}(\gamma C(s_{t+1},a_{t+1}))$

Here the action of next state here is based on the max of $\mathrm{Q}^{\mathrm{combined}}(s_{t+1},a_{t+1})$

As mentioned in 5.b, curiosity reward will definitely contributes in the error of value estimation, but it doesn't seem to cause a big problem in the DQN-Curiosity papers I have seen. From this observation, can the approximation of the state action value of intrinsic reward at s_t not follow the bellman equation and we still achieve a good results?

One experiment I would like to try is to approximate C(s_t, a_t) to the multiplication of Novelty(s_{t+1}) and max novelty of states of the same episode after s_t.