Adversarial Inverse Reinforcement Learning in Dynamic Environment

Samin Yeaser Arnob
ID: 260800927

COMP-767-Reinforcement Learning,
McGill University

Goal of the Project:

- Compare Imitation and Inverse Reinforcement Learning in dynamic environment
- Look for at what extend Reward function help IRL when dynamic changes
- Explore better reward function for IRL

Dynamic Environments:

- Minigrid
- DeepLab
- Mujoco

Completed Tasks:

- Implemented Imitation learning
- Compare performance for different policy learning algorithm
- Performance in changing dynamics:
- Crippled front legs
- Added Noise in action

Performance Evaluation of TD3 and SAC policy

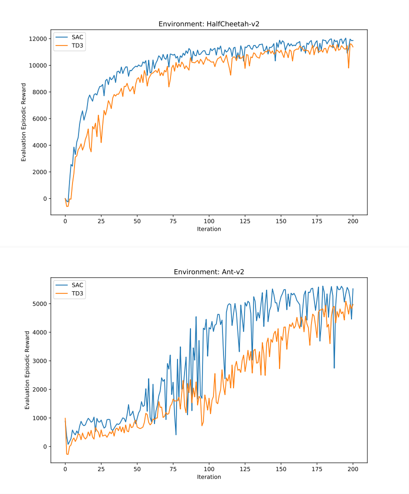

Figure 1: Policy Evaluation in RL setting

Performance Evaluation of TD3 and SAC policy

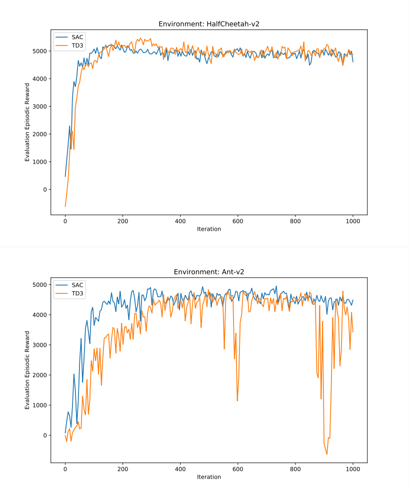

Figure 2: Policy Evaluation in Imitation Learning setting

Performance

Performance in Changed dynamics:

- Imitation Learning setting
- Policy used: Soft-Actor-Critic (SAC)
- Experiment was done on Ant-v2

Policy	TD3	SAC
Crippled leg	-2322847	-2322884
TD Noise	1941.51	977
Gaussian Noise	660.38	854
Actual Performance	3721	4482

Maximum entropy IRL

Objective: Want reward function that maximizes likelihood of (expert) trajectories $\theta = argmax_\theta \log \prod_{\tau_d \in D} P(\tau_d) = \sum_{\tau_d \in D} \log P(\tau_d)$ Now,

$$
\begin{aligned}
L =& \frac{1}{M} \sum_{\tau_d \in D} \log \frac{1}{Z} e^{-c(\tau_d)} \\
=& \frac{1}{M} \sum_{\tau_d \in D} \log e^{-c(\tau_d)} + logZ
\end{aligned}
$$

Where Z = partition function = $\sum_\tau e^{-c(\tau)}$ sum over/ integration over all possible trajectory

* $\nabla_\theta L = \frac{1}{M} \sum_{\tau_d \in D} d\frac{c(\tau)}{d\theta} + \sum_s p(s|\theta, T) * d\frac{c(s)}{d\theta}$

Algorithm:

* $\pi(a|s) \rightarrow \mu(s) \rightarrow P(s|\pi(\theta)) \rightarrow \nabla L(\theta)$ *

$\mu(s) = \sum_a \sum_s \mu(s) * \pi(a|s) * \underbrace{p(s'|s, a)}$

GAN-GCL

$$D = \frac{1/z \times e^{-c(\tau)}}{1/z \times e^{-c(\tau)} + q(\tau)}$$

$$D = \frac{p(\tau)}{p(\tau) + q(\tau)}$$

Porblem with GAN-GCL

$$D_\theta(\tau) = \frac{exp f_\theta(\tau)}{exp f_\theta(\tau) + \pi(\tau)}$$

Considering full-trajectory result in high variance.

AIRL

Instead of considering trajectory use every (s, a) pairs that reduces the variance

$$D_\theta = \frac{expf_\theta(s, a)}{expf_\theta(s, a) + \pi(s, a)} \tag{1}$$

Distangled reward:

Reward function is distanlged when under all dynamics optimal policy is same $\pi^*_{r', T}(a|s) = \pi^*_{r,T}(a|s)$

For transition dynamic $T(s, a) = s'$ reward function can be written as:

$$\hat{r}(s, a) = r(s, a) + \gamma\phi(s') - \phi(s)$$
$$\hat{r}(s, a) = r(s, a) + \gamma\phi(T(s, a)) - \phi(s)$$

For two different MDP, M and M' ; if the transition dynamics are T and T'

AIRL

NOTE: (To remove unwanted reward shaping) Learned reward function can only depend on the current state, s

$$D_\theta = \frac{expf_\theta(s,a)}{expf_\theta(s,a) + \pi(s,a)}$$

$$D_{\theta,\phi} = \frac{expf_{\theta,\phi}(s,a)}{expf_{\theta,\phi}(s,a) + \pi(a|s)}$$

Where, reward approximator $g(\theta)$ and shaping term h_ϕ

$$f_{\theta,\phi} = \underbrace{g_\theta(s,a)}_{reward\ \ approximator} + \gamma h_\phi(s') - h_\phi(s)$$

Update function:

$$r_{\theta,\phi}(s,a,s') \leftarrow \log D_{\theta,\phi}(s,a,s') - \log(1 - D_{\theta,\phi}(s,a,s')) \qquad (2)$$

Future Work:

- Performance of Inverse Reinforcement Learning in dynamic MuJoCo environments
- Will evaluate performance in Maze
- Will evaluate performance DeepLab

References

1. Ziebart, B., Maas, A., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In AAAI Conference on Artificial Intelligence, 2008.
2. Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. abs/1611.03852, 2016a.
3. J. Ho and S. Ermon, "Generative adversarial imitation learning," in Advances in Neural Information Processing Systems, pp. 4565–4573, 2016.
4. Fujimoto, S., van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. arXiv preprint arXiv:1802.09477, 2018.
5. T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv preprint arXiv:1801.01290, 2018.

References

6. Fu, J., Luo, K., and Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. In International Conference on Learning Representations (ICLR)

7. Andrew Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In International Conference on Machine Learning (ICML), 1999.