1 Summary

1. Goal-reaching practice is an effective form of self-supervised learning for robotics.
2. But defining a reward fun. in high dimensional space is hard.
3. The breath first nature of existing strategy (HER) means that some areas of the state space takes a long time to learn.
4. They propose goalGAIL, which can boost goal-conditioned policy learning with only state demonstrations.

2 Demonstrations in goal-conditioned tasks

2.1 Goal-conditioned behavior cloning

1. Assume the demo is $\left\{\left(s_0^j, a_0^j, s_1^j, \ldots\right)\right\}_{j=0}^D$, the BC loss is:

$$\mathcal{L}_{\mathrm{BC}}(\theta, \mathcal{D}) = \mathbb{E}_{\left(s_t^j, a_t^j, g^j\right) \sim \mathcal{D}}\left[\left\|\pi_\theta\left(s_t^j, g^j\right) - a_t^j\right\|_2^2\right]$$

2. This loss can be combined with gradient descent in other policy loss, for example the DDPG loss:

$$\nabla_\theta \hat{J} = \frac{1}{N}\sum_{i=1}^{N} \nabla_a Q_\phi(a, s, g) \nabla_\theta \pi_\theta(s, g)$$

3. The improvement guarantees wrt the task reward are lost when we combine the BC and the deterministic policy gradient updates, but this can be sidestepped by annealing the BC loss.

2.2 Relabeling the expert

1. The expert trajectories have been collected by asking the expert to reach a specific goal g^i.
2. If we have the transition $\left(s_t^j, a_t^j, s_{t+1}^j, g^j\right)$ in a demo, then we can also consider the transition $\left(s_t^j, a_t^j, s_{t+1}^j, g' = s_{t+k}^j\right)$ as a transition when the expert is trying to reach goal g'.
3. They assume that the task is quasi-static.

2.3 Goal-conditioned GAIL with hindsight

1. GAIL can leverage rollouts collected by the learning agent in an unsupervised manner.
2. They extend GAIL to tackle goal-conditioned tasks, and then describe how it can be combine with HER.
3. The discriminator conditions on the goal $D_\psi(a, s, g)$ and can be trained to minimize:

$$\begin{aligned}\mathcal{L}_{GAIL}\left(D_\psi, \mathcal{D}, \mathcal{R}\right) =& \mathbb{E}_{(s,a,g)\sim\mathcal{R}}\left[\log D_\psi(a, s, g)\right] + \\ & \mathbb{E}_{(s,a,g)\sim\mathcal{D}}\left[\log\left(1 - D_\psi(a, s, g)\right)\right]\end{aligned}$$

4. Once the discriminator is trained, $\log D_\psi\left(a_t^h, s_t^h, g^h\right)$ is used as the reward function to train an RL agent.
5. They combine this loss with DDPG and indicator reward function $r_t^h = 1\left[s_{t+1}^h == g^h\right]$.

2.4 Use of state-only demonstration

1. They replace the actions in the GAIL formulation by the next state s'.
2. Given a desired goal, they argue that it should be possible to determine if a transition s, s' is taking the agent in the right direction.
3. The new GAIL loss is:

$$\mathcal{L}_{GAIL^s}\left(D_\psi^s, \mathcal{D}, \mathcal{R}\right) = \mathbb{E}_{(s,s',g)\sim\mathcal{R}}\left[\log D_\psi^s\left(s, s', g\right)\right] + \mathbb{E}_{(s,s',g)\sim\mathcal{D}}\left[\log\left(1 - D_\psi^s\left(s, s', g\right)\right)\right]$$

3 Experiments

1. The goal and state space in all task are low-dimensional and not vision-based.
2. They show that goalGAIL outperforms HER and GAIL.
3. HER has slow convergence, though it reaches the same final performance if run for long enough.
4. GAIL learns fast at the beginning, but its final performance is capped.
5. They argue that GAIL has this behavior because despite collecting more samples, those come with no reward indicating what the task is. Therefore, once the discriminator has extracted all the info from the demonstrations, it can not keep learning any generalize to goals further from the demo.
6. They argue that their expert relabeling techniques help to overcome the shortcoming of GAIL.
7. They show that expert relabeling helps for BC, HER + BC and goalGAIL.