1 Summary

1. They show that a critical challenge in applying adversarial imitation from high-dimensional sensory data is the tendency of the discriminator network to use task-irrelevant features in the state to discriminate between the expert and imitator trajectories.

2. They propose to train the discriminator with a modified obj, such that the discriminator is penalized if it uses task-irrelevant feature to discriminate between states generated by the expert and states generated by the imitator.

2 Background Info on Adversarial Imitation

1. GAIL is used to derive a reward function from expert demonstration.

2. In GAIL, a discriminator network $D(s, a)$ is trained to distinguish between expert and imitator state-action pair.

3. The GAIL obj is:

$$\min_{\pi} \max_{D} \mathbb{E}_{(s,a)\sim\pi_E}[\log D(s, a)] + \mathbb{E}_{(s,a)\sim\pi}[\log(1 - D(s, a))] - \lambda_H H(\pi)$$

   where π_E is the expert policy, π is the imitator, and H is an entropy term.

4. The reward function to train π is: $R(s, a) = -\log(1 - D(s, a))$.

3 Task-relevant Adversarial Imitation Learning

1. They argue that the discriminator can focus on task-irrelevant feature of the observations while ignoring the differences in the behavior of the expert and imitator (I take behavior here to mean the induced state visitation dist).

2. If this happens, then the reward function derived from the discriminator is not informative to train the imitator.

3. They argue that the discriminator should only be able to distinguish between states generated by expert and imitator where there are meaningful behavior differences.

4. They thus propose TRAIL, which trains the parameter ψ of the discriminator by solving the optimization problem:

$$\begin{array}{ll} \max_{\psi} & \mathbb{E}_{s\sim\pi_E}\left[\log D_\psi(s)\right] + \mathbb{E}_{s\sim\pi_\theta}\left[\log\left(1 - D_\psi(s)\right)\right] \\ \text{s.t.} & \frac{1}{2}\mathbb{E}_{s\sim\pi_E}\left[\mathbf{1}_{\mathrm{D}_\psi(\mathrm{s})\geq\frac{1}{2}} | s \in \mathcal{I}\right] + \frac{1}{2}\mathbb{E}_{s\sim\pi_\theta}\left[\mathbf{1}_{\mathrm{D}_\psi(\mathrm{s})<\frac{1}{2}} | s \in \mathcal{I}\right] \leq \frac{1}{2} \end{array}$$

   where $\mathcal{I}$ is called the invariant set.

5. The constraint of the opt problem states that for state in the invariant set, the discriminator should not be able to distinguish between states generated by the expert and states generated by the imitator.

6. The idea is thus, the state in the invariant set should contain task-irrelevant features. By training with this obj, the discriminator is trained to ignore the task-irrelevant features.

4 The selection of the invariant set

1. They advocate for using early frames from the both expert and imitator episodes as the invariant set.

2. They argue that in early frames, there is little to no task-specific behavior exhibited by either and only task-irrelevant feature is present.

5 Interesting tidbit from experimental sections

1. Data augmentation helps with reducing discriminator overfitting.

2. Actor early stopping: when the imitator has learnt the desired behavior, and the resulting data is used to train the discriminator, the discriminator will be unable to distinguish between expert and imitator states based only on task-specific feature. The discriminator is thus forced to rely on task-irrelevant feature. To avoid this, they devise a heuristics so that successful behavior by the imitator is rarely present in the training data of the discriminator.

3. Figure 3 lists actor step as the label of the x axis. If I take this to be the number of environment interaction, then the number of environment interaction required is still massive.

6 Questions

1. What does it mean for the discriminator to overfit? What is it overfitting to and what is it supposed to generalize to?

2. It is unclear what is happening in the lift distracted seeded experiment.