Démarche

Mehdi Mounsif

19 mars 2018

1 Récap : The day before

- Lecture de [1] et [3]. GAIL semble être un voie plus prometteuse que Guided Cost Learning, notamment parce que GAIL permet d'obtenir la politique et non pas seulement la fonction de récompense. A relire.
- Préparation DDPG
- GAN : Tests avec WGAN sur dataset pokémon et robots. Pas exceptionnel. Essayer [2]

2 Résultats de la réunion

Intégrer plus de robotique. Permettra de publier rapidos (et d'obtenir plus de liberté).

Feuille de route : Robotique

- Modèle géométrique indirect. Fonctionne. Transposée de la Jacobienne itérative.
- Déplacer le robot avec ces modèles analytiques. Fonctionne sur LowReacher et Reacher
- Injecter du bruit dans les capteurs et le gérer avec RL.

Feuille de route : IA

- DDPG
- IRL : Regarder GAIL [3]
- GAN : Des progrès. MNIST. Génération Pokémons (approximative)
- Model-based : Pas étudié.

Feuille de route : Misc

- Tester PPO sur Atari : Implémentation propre de AC, A2C, PPO à la maison. Rajouter CNN et tester sur Atari
- Tester sur SNESx9 : Détection de rectangles avec un réseau de neurones.
- Animation IK : Pas de progrès

3 DDPG

Implémentation d'une version allégée de DPPG. Comprend :

- Acteur, critique
- Target networks
- OU noise process

Manque Experience Replay. En tout cas, les résultats sont nuls. Ajouter Experience Replay et vérifier à nouveau.

4 GAIL : Generative Adversarial Imitation Learning

Permet d'apprendre une politique IRL-style sans s'encombrer de la fonction de récompense [3]. Semble, à priori, plus approprié que [1].

En pratique, les auteurs introduisent une fonction de régularisation ψ qui cherche à minimiser la distance entre ce qu'ils appellent **occupancy measure** de l'expert et celle de la politique en cours d'apprentissage. Cette notion représente la distribution des paires état-actions rencontrés suivant la politique.

On a :

$$\psi_{GA} = \begin{cases} \mathbb{E}_{\pi_E}[g(c(s,a))] \text{ si } c < 0 \\ +\infty \text{ sinon} \end{cases} \qquad where \quad g(x) = \begin{cases} -x - \log(1 - e^x) \text{ si } x < 0 \\ +\infty \text{ sinon} \end{cases} \tag{1}$$

J'ai découvert des approches plus récentes :

- Robust Imitation of Diverse Behaviours [6]
- Third Person Learning [5]
- Ainsi qu'une review : [4]

Références

[1] Finn, C., Levine, S., and Abbeel, P. Guided cost learning : Deep inverse optimal control via policy optimization. *CoRR abs/1603.00448* (2016).

[2] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved Training of Wasserstein GANs. *ArXiv e-prints* (Mar. 2017).

[3] Ho, J., and Ermon, S. Generative adversarial imitation learning. *CoRR abs/1606.03476* (2016).

[4] Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. Imitation learning : A survey of learning methods. *ACM Comput. Surv. 50*, 2 (Apr. 2017), 21 :1–21 :35.

[5] Stadie, B. C., Abbeel, P., and Sutskever, I. Third-Person Imitation Learning. *ArXiv e-prints* (Mar. 2017).

[6] Wang, Z., Merel, J., Reed, S., Wayne, G., de Freitas, N., and Heess, N. Robust Imitation of Diverse Behaviors. *ArXiv e-prints* (July 2017).