

Démarche

Mehdi Mounsif

20 mars 2018

1 Récap : The day before

- Articles sur imitation : Third Person Imitation [4] et [5].
- Implémentation complète DDPG (target networks, experience replay et OU noise) . Résultats médiocres. A revoir.
- PPO en test : Sélection systématique de l'action ayant la plus haute probabilité. Ne tremble pas. Aurait besoin d'un peu de finesse supplémentaire pour augmenter ses performances.
- Question : Comment choisir une action continue avec PPO ?

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 [2]
- 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 [2]. 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 :

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

J'ai découvert des approches plus récentes :

- Robust Imitation of Diverse Behaviours [5]
- Third Person Learning [4]
- Ainsi qu'une review : [3]

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] HO, J., AND ERMON, S. Generative adversarial imitation learning. *CoRR abs/1606.03476* (2016).
- [3] 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.
- [4] STADIE, B. C., ABBEEL, P., AND SUTSKEVER, I. Third-Person Imitation Learning. *ArXiv e-prints* (Mar. 2017).
- [5] WANG, Z., MEREL, J., REED, S., WAYNE, G., DE FREITAS, N., AND HEES, N. Robust Imitation of Diverse Behaviors. *ArXiv e-prints* (July 2017).