Démarche

Mehdi Mounsif

22 mars 2018

1 Récap : The day before

- Articles sur imitation : Third Person Imitation [6] et [7].
- 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 DDPG

Toujours rien du côté de DDPG. Malgré une nouvelle approche d'implémentation, les résultats sont toujours à pleurer. Pas de commentaires particuliers. A revoir.

3 PPO continuous

Important : Dans PyTorch, copier les données d'une Variable peut avoir des conséquences désastreuses. Je soupconne l'architecture de copier directement l'adresse du pointeur. Pourrait expliquer le problème du ratio constamment égal à 1. Préférer l'utilisation de `state_dict` et `load_state_dict`.

Pour l'aspect continu : D'après une implémentation du forum PyTorch (parce que ce n'est pas clairement expliqué dans [5] peut-être dans [4]), il est possible d'avoir en sortie de π deux têtes :

- Une pour la moyenne
- Une pour l'écart-type

On génère alors une action issue de cette distribution ie : $a \sim \mathcal{N}(\mu, \sigma)$ et on prend la probabilité log pour la backprop. L'implémentation est sur Pendulum et semble fonctionner. Quid de plusieurs actions ?

4 IK

Modification de ReacherIK. Ajout du solver de jacobienne. L'idée est que l'agent choisisse une action parmi n (dans ce cas, 8, correspodant aux directions cardinales) et que le vecteur soit transmis à la jacobienne pour qu'elle calcule l'incrément dans les angles.

Question : Serait-il possible d'approximer des jacobiennes plus importantes pour robotique ?

5 FloydHub

Découverte du service FloydHub pour location d'architecture pour Deep Learning. 12 dollars pour 10 heures de GPU. Parfait pour les GANs.

6 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

7 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.

8 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 :

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

J'ai découvert des approches plus récentes :

- Robust Imitation of Diverse Behaviours [7]
- Third Person Learning [6]
- 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] Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. Trust region policy optimization. *arXiv* (April 2017).

[5] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR* (2017).

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

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