Démarche

Mehdi Mounsif

1er mars 2018

1 Récap : The day before

- PPO : J'ai observé que le ratio était constamment égal à 1 sur Reacher. En revanche, sur Cartpole il est différent à chaque itération. Quel est le problème ?
- Lectures sur GANs. Découverte de l'architecture DCGAN et de la problématique de l'approximation de distribution (différents résultats si MaxLikelihood ou Jensen-Shannon)
- Idée RL : HER - tirer des leçons des échecs (mise en abîme ?). Voir [2]
- Malgré les déclarations, j'ai l'impression que les approches alternatives sont loin d'être prêtes à rivaliser avec les réseaux de neurones (NuPIC est incompréhensible, Vicarious n'a rien fait depuis 2013, Neurithmic n'est pas vraiment à jour non plus).

2 PPO avec convolution

Dans [1], on trouve un récapitulatif des architectures utilisées en RL. On y trouve notamment la structure et les paramètres des couches de convolution. Ainsi, dans [5] pour le benchmark Atari, les auteurs utilisent :

- (Conv2D - 16 channels - stride 2)x2
- Fully connected avec 20 neurones

Les résultats de TRPO sur Atari n'étant pas cosmiques, je choisis l'architecture de [6] :

- Conv2D 16 filtres de $8x8$ avec stride $= 4$
- Conv2D 32 filtres de $4x4$ avec stride $= 2$
- FC 256
- NL : ReLU

3 PPO sans convolution

Test de l'approche de [3]. Au lieu de :

1. Inputs $\rightarrow$ Couches cachées $\rightarrow$ Softmax
2. Categorical(probs) $\rightarrow$ Action

Il propose :

1. Inputs $\rightarrow$ Couches cachées $\rightarrow$ Categorical
2. Sample distrib $\rightarrow$ Actions & log probability.

Ce n'est pas un grand changement, mais peut éventuellement me mettre sur la piste de l'erreur. Penser à faire plotter pour la state value

4 IRL - GANs

Essayer de trouver un exemple de GANs

5 DDPG

A faire demain

6 HER - [2]

L'idée découle de UVFA (Universal Value Function Approximator) [4], une extension de DQN où on considère une configuration dans laquelle plusieurs buts/objectifs sont possibles. A chaque début d'épisode, on considère la paire (s_i, g_i) correspondant à l'état initial et au but de cet épisode. Avec cette plate-forme, la fonction de récompense dépend classiquement de la paire (s_t, a_t) mais également de g_i.

Pour HER, on propose de délaisser les *shaped rewards*, qui sont tributaire d'une connaissance de l'environnement. Pour les remplacer et conserver les performances malgré les récompenses éparses, considérons ceci : Soit un épisode avec une trajectoire d'états $\zeta = \{s_0, s_1, ..., s_n\}$ et un but $g \neq s_1, ...s_n$. Par conséquent, l'agent à reçu une récompense négative à chaque pas de temps. Cette trajectoire ne nous donne aucune information sur comment atteindre le but g mais en revanche apporte une solution pour atteindre s_t. Donc, si en off-policy on remplace g par s_t, on peut considérer des trajectoires gagnantes ie : qui contiennent de l'information pertinente.

7 RL

Tester si possible les nouvelles plateformes de Gym. Résultat : Impossible, fonctionne avec MuJoCo.

J'ai créée un outil de visualisation des states values qui a permis de confirmer l'apprentissage de la value function. De plus, un autre outil, appelé Planner a permis de complexifier de manière croissante la tâche (notamment en ajoutant des cibles graduellement). On obtient des performances plus consistantes. La classe est disponible sur GitHub.

Références

[1] 5vision. Deep reinforcement learning networks. https ://github.com/5vision/deep-reinforcement-learning-networks, 2016.

[2] Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. Hindsight Experience Replay. *ArXiv e-prints* (jul 2017).

[3] Kostrikov, I. Pytorch implementations of reinforcement learning algorithms. https ://github.com/ikostrikov/pytorch-a2c-ppo-acktr, 2018.

[4] Schaul, T., Horgan, D., Gregor, K., and Silver, D. Universal value function approximators. In *ICML* (2015).

[5] Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. Trust region policy optimization. *arXiv* (April 2017).

[6] Volodymyr, M., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Daan, W., and Martin, R. Play atari with deep reinforcement learning. Tech. rep., 2013.