Démarche

Mehdi Mounsif

22 février 2018

J'ai utilisé le système FastPlot pour observer un agent lors du processus d'apprentissage. Certain points sont à améliorer, notamment l'utilisation d'une colormap pour les poids, mais c'est globalement utile. Pour plus de confort, peut-être une version PyQt4 serait pertinente. Elle permettrait de switcher entre les différentes fenêtres puisque l'affichage des poids est encombré. La visualisation des décisions est en revanche satisfaisante .

Poursuite de PPO et étude de DPPG.

- Pour PPO : Création d'une fonction sensée jouer le rôle de l'ancienne politique. Elle évaluera les actions choisies et permettra de calculer le ratio
- Pour DDPG : Le système de mémoire n'est pas entièrement satisfaisant. Pour sampler un batch pour l'entrainement, j'utilise random.sample après avoir shufflé la mémoire. Pour finir, une fois le batch obtenu, je supprime les éléments choisis de la mémoire. En théorie, le problème est qu'il est possible que des expériences dépassés soient conservées et utilisées plus tard. En pratique, je ne sais pas quel impact cela peut avoir.

J'ai remis à jour le système d'enregistrement des trajectoires générées par un expert (moi) controllant le robot pour atteindre des cibles. On pourrait alors s'en servir pour :

- Inverse RL
- LSTM
- Régression classique

J'ai crée un controlleur qui permet de manipuler le robot. Ce controlleur m'a permit de créer quelques dizaines de trajectoires manuellement. Du point de vue quantitatif, il serait plus rentable d'implémenter une pseudo-inverse mais la création manuelle permet d'enregistrer les actions, là où la pseudo-inverse donnerait lieu à l'enregistrement des couples et des vitesses angulaires. Souhaitable dans le cas de DDPG. J'ai généré plusieurs trajectoires que j'ai ensuite approximées avec un réseau d'architecture suivante :

- Observations - ReLU - 256
- 256 - ReLU - 128
- 128 - Softmax - Actions | 128 - Actions

Optimisation avec ADAM, avec $\alpha = 10^{-4}$. Différentes tailles de batch ont été testées (10 - 32 - 64). L'approche est peu fructueuse. Globalement, même après 100000 itérations, le comportement de l'agent est loin de celui de l'expert. Par conséquent :

- Pourrait éventuellement servir de point de départ pour RL
- Chercher à prédire les états plutôt que les actions. Ceci implique de ne pas shuffler le dataset (?)
- Approche Inverse learning avec entropie.

Références