Démarche

Mehdi Mounsif

2 mars 2018

1 Récap : The day before

- Création de deux outils qui ont permis une amélioration des performances dans LowReacher. Je ne sais pas si PPO est une clé de voûte de cette amélioration (i.e : pourrait-elle fonctionner avec A2C ?). Je n'ai pas réglé le problème du ratio. Voir ce que donne l'implémentation [2]
  - FastValue : Visualisation de la state value en temps réel. A permis de voir que le critique mesurait (j'ai l'impression) avec une cohérence grandissante les états au cours du temps.
  - Planner : Mise en place d'un curriculum pour l'agent. Amélioration drastique de la performance.
- Lectures de l'article Hindsight Experience Replay [1]. Bonne idée, mais en pratique comment remplacer les buts dans les trajectoires ? Si les Variables de PyTorch contiennent les informations du *flow* alors elles sont liées aux états qui les ont générées.
- Les nouveaux environnements de Gym dépendent de MuJoCo. Peu pratique. Tenter d'installer MuJoCo.
- J'ai proposé une réunion pour la semaine prochaine.

2 PPO : Nouvelle implémentation

Calcul de l'entropie :

$$H = -\sum_i \log(p(x_i)) * p(x_i)$$

A utiliser dans la fonction de coût.

- Pour A2C :

$$\mathcal{L}_{total} = alpha\mathcal{L}_{value} + \mathcal{L}_{policy} - \beta\mathbb{E}[H]$$

  - ☐ Avec $\mathcal{L}_{value} = TD_{error}$
  - ☐ $\mathcal{L}_{policy} = -\log(\pi(a|s,\theta)) * \mathbb{A}$ où $\mathbb{A}$ est l'avantage
- Pour PPO :

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{value} + \mathcal{L}_{policyPPO} - \beta\mathbb{E}[H]$$

3 DDPG

4 GANs

- Biblio GANs
- Biblio GANs en robotique

Références

[1] 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).

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