Démarche

Mehdi Mounsif

7 mars 2018

1 Récap : The day before

- Lecture de [2]. Des bonnes idées, mais quelques points non clairs. J'ai envoyé un mail à l'auteur principal qui m'a répondu. Figure 1 est le schéma de l'architecture de la fonction
- Création d'un dataset et début de l'implémentation d'un GAN. A finir aujourd'hui

2 Learning without a reward function – Suite

Dans [2], les auteurs proposent une méthode de RL qui permet d'apprendre en se détachant de la fonction de récompense. La méthode fait appel au principe d'entropie maximum et à la théorie de l'information. La brique de base est dénommée *skill*. Chaque *skill*est en fait une politique π qui doit être aussi différent que possible des autres *skills*(dans le sens où les états visités sous l'influence de ce *skill*sont les plus différents possibles que ceux visités avec un autre *skill*).

2.1 Principe

Deux phases :

- Exploration non-supervisée : Exploration de l'espace, pas de récompense.
- Stage supervisé : RL classique. L'agent reçoit une récompense et doit la maximiser.

Fonctionnement n'est pas sans évoquer [1]. En effet, on peut comparer la partie non-supervisée à du peaufinage de talent(farming). Les points clés sont :

- Les *skills*doivent se distinguer au niveau des états (au plutôt, on doit pouvoir distinguer les *skills*)
- L'exploration est encouragée en maximisant l'entropie entre les *skills*

Formellement, on chercher à maximiser l'information entre les *skills*et les états $MI(s, z)$, où z correspond à un *skill*. Egalement, on minimise l'information entre les *skills*et les actions en fonction de l'état $MI(a, z|s)$. On veut pour finir maximiser $\mathcal{H}(a|s)$. Soit :

$$\mathcal{F}(\theta) = MI(s, z) + \mathcal{H}(a|s) - MI(a, z|s)$$

Que l'on peut écrire :

$$\mathcal{F}(\theta) = \mathcal{H}(a, z|s)\mathcal{H}(z) - \mathcal{H}(z|s)$$

A finir. J'ai envoyé un mail à Ben Eysenbach `eysenbachbe@google.com` (GoogleBrains) pour demander des clarifications sur les *skills*.

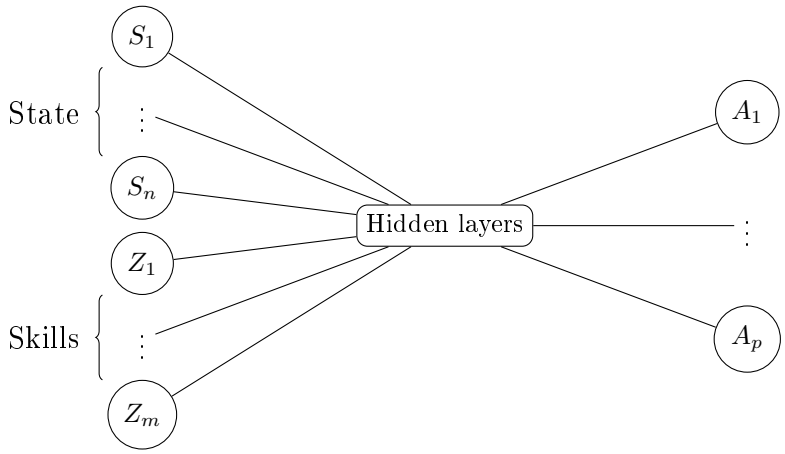

FIGURE 1 – DIYAN - Input space representation

3 GANs

Fin de l'implémentation et visualisation des résultats. Peu probants. Il faudrait introduire la distance de Watterstein comme fonction de coût. Vérifier implémentation suivant : `https://github.com/t-vi/pytorch-tvmisc/blob/master/wasserstein-distance/Improved_Training_of_Wasserstein_GAN.ipynb`

4 DPPG - PPO

Toujours pas de progrès

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] EYSENBACH, B., GUPTA, A., IBARZ, J., AND LEVINE, S. Diversity is All You Need : Learning Skills without a Reward Function. *ArXiv e-prints* (Feb. 2018).