Projet DAC

ADDAD Youva

Sorbonne Université

2021

Table des matières

Introduction

Les satellites d'observation de la Terre sont des senseurs qui acquièrent des données, les compressent et les mémorisent à bord, puis les vident vers des stations.

Avec l'augmentation du nombre de satellites d'observation de la Terre planifier les activités de vidage du satellite est de plus en plus problématique.

Nous allons donc appliquer des modèles de Deep Reinforcement Learning afin d'optimiser le plan de vidage.

Difficulté du problème

$$\begin{array}{lll} \min & \textit{Perte} & \\ \text{s.t.} & \min & \textit{Connexion} \\ & \min & \textit{expiration} \\ & \max & \textit{stockage} \end{array} \tag{1}$$

Ce probléme est un probléme NP-Hard

Garey, Michael R., et David S. Johnson. Computers and Intractability ; A Guide to the Theory of NP-Completeness. W. H. Freeman Co., 1990.

Deep Reinforcment Learning

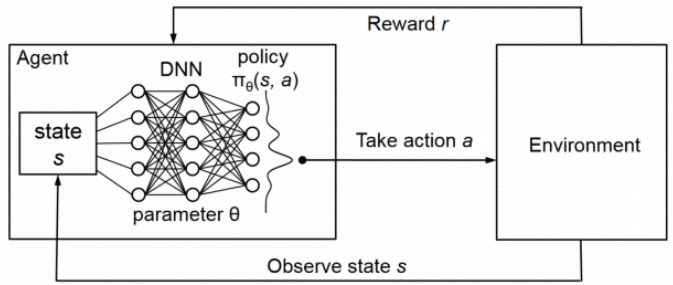

Figure – Une Architecture typique de Deep Reinforcment Learning -DQN

-Un réseau neuronal profond est caractérisé par une succession de traitement des couches.
-Chaque couche consiste en une transformation non linéaire et la séquence de ces transformations conduit à apprendre différents niveaux d'abstraction.

Fonction Objective

$$Q^{new}(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\Big(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} - \underbrace{Q(s_t, a_t)}_{\text{old value}}\Big)}^{\text{temporal difference}} \quad (2)$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{new value (temporal difference target)}}$$

Fonction objective

Avec r_t la récompense reçue lors de la transition vers l'état s_t, α est le taux d'apprentissage ($0 < \alpha \leq 1$).
Donc de manière général $Q^{new}(s_t, a_t)$ est la somme de :

- $(1 - \alpha).Q(s_t, a_t)$ la valeur actuelle pondérée par le taux d'apprentissage. Les valeurs du taux d'apprentissage proches de 1 accélèrent les changements de Q.
- $\alpha.r_t$ la récompense a obtenir si l'action a_t a été entreprise dans l'état s_t pondéré par le taux d'apprentissage.
- $\alpha.\gamma \max_a Q(s_{t+1}, a)$ la récompense maximale qui peut être obtenue de l'état pondéré par le taux d'apprentissage.

OpenAI gym

Pour pouvoir appliqué un modèle de DRL et optimisé la fonction objective précédente nous devons adapter notre modèle a openAI gym.

- Gym est un toolkit pour développer et comparer des algorithmes d'apprentissage par renforcement.
- Le noyau de l'interface de gym est env, Les méthodes de env :
  - reset(self) : Réinitialisez l'état de l'environnement. Renvoie une observation.
  - step(self, action) :Marche de l'environnement d'un pas de temps. Renvoie une observation, récompense, fait(done), info.
  - render(self, mode='human') : Eend une image de l'environnement.

Modélisation

Modélisation d'un Simulateur pour appliquer l'apprentissage par renforcement.

Satellite

Un satellite est représenté par :

- Son déplacement **Position,direction ,vitesse**
- Sa mémoire : **Mémoire, Mémoire libre, taux de stockage réel, taux de stockage estimé sur ces prochaines acquisitions**.
- Les tâches à faire
- Les observations stockées en mémoire.
- Une maille.

Station

Une Station est représentée par :

- Sa Position
- Son débit de transmission de données.
- Son état (libre , connectée)
- Un temps de connexion
- Le nombre de satellites pour les quels elle est visible.

Tache , Image

- **Les taches** sont modélisées par une position , une longueur , une validité.
- **Les images** sont crée suite à une acquisition d'une tâche sont modélisées par une validité et une taille .

Modèles

Nous avons crée deux modèles d'apprentissage conformes à l'environnement GYM. Un Modèles est un environnement qui est initialisé avec une liste de satellites et une liste de stations, toute action de déchargement possible par un des satellites est une nouvelle étape d'apprentissage.

Multi selection

Ce modèle est basé sur le choix d'une politique de déchargement en fonction de l'état du satellite et de la station concernés

- Observation : c'est l'état du satellite courant dans le quel on trouve les caractéristique du satellites a l'instant (état mémoire, estimation des nouvelles acquisitions) et les statistique sur les attributs **priorité , size, validité** des acquisitions sauvegardé en mémoire , des acquisitions possible a décharger , des prochaines taches.
- Les Actions Possibles
  - Ne pas se connecter
  - Se connecter et décharger par maximum de priorité.
  - Se connecter et décharger par maximum de taille.
  - Se connecter et décharger par minimum de validité.

Binary Model

Ce modèle est basé sur la décision

- Observation :Contient l'observation du modèle MultiChoix en lui ajoutant les caractéristique d'une acquisition choisis par maximum de priorité parmi celles qui peuvent tre déchargées.
- Les Actions Possibles
  - Ne pas se connecter
  - Se connecter et décharger le choix présenté dans l'observation.

Récompense

La récompense est commune pour les deux modèles :

- Toutes acquisitions déchargées nous recomposent par sa priorité.
- Chaque connexion a une station coûte le temps de connexion.
- Chaque acquisition sauvegardée expirée nous coûte sa priorité.
- Chaque tâche qu'on n'a pas pu réaliser dû à une saturation mémoire, nous coûte sa priorité.

Expérience

Expérience

Évaluation

Évaluation

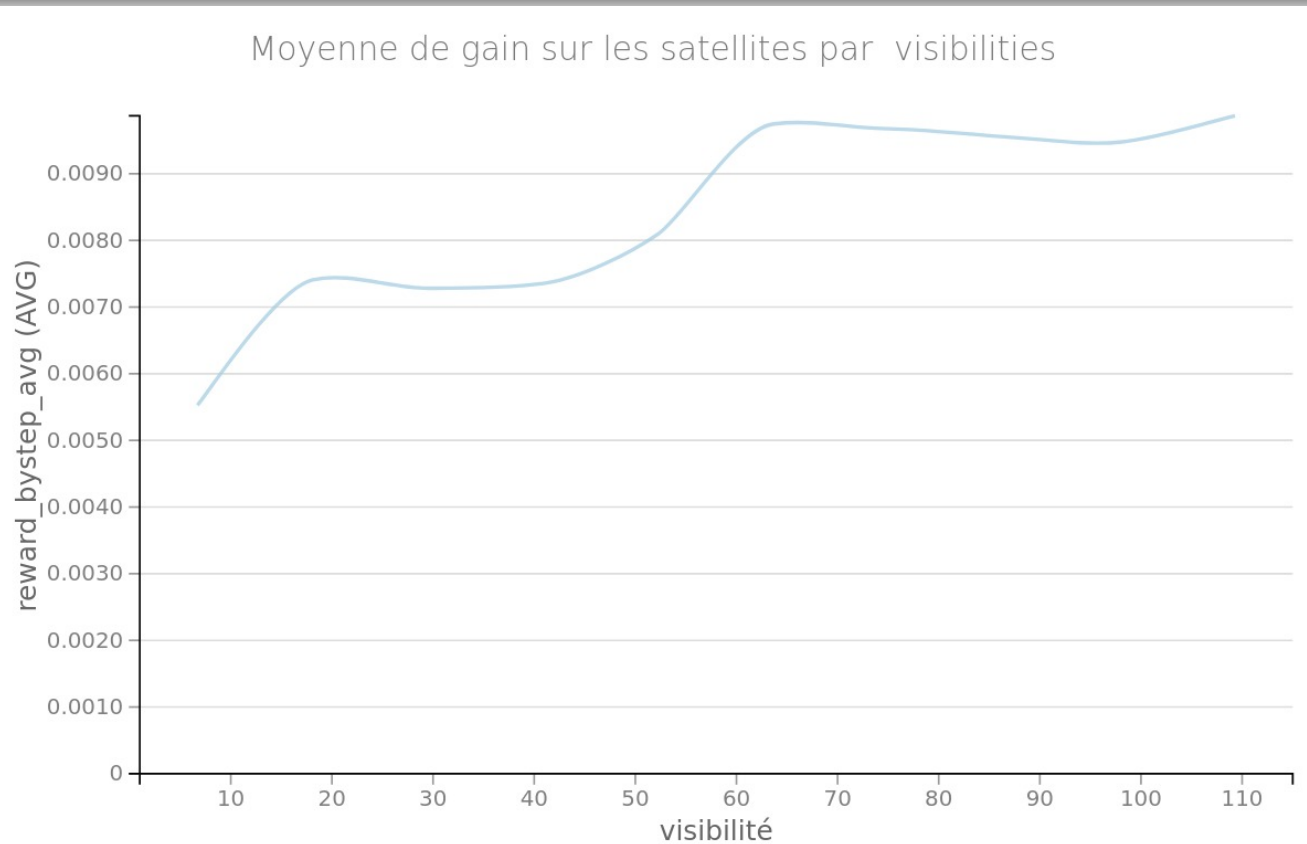

Figure – Courbe de la moyenne du gain sur tout les satellites en fonction du nombre de visibilités station

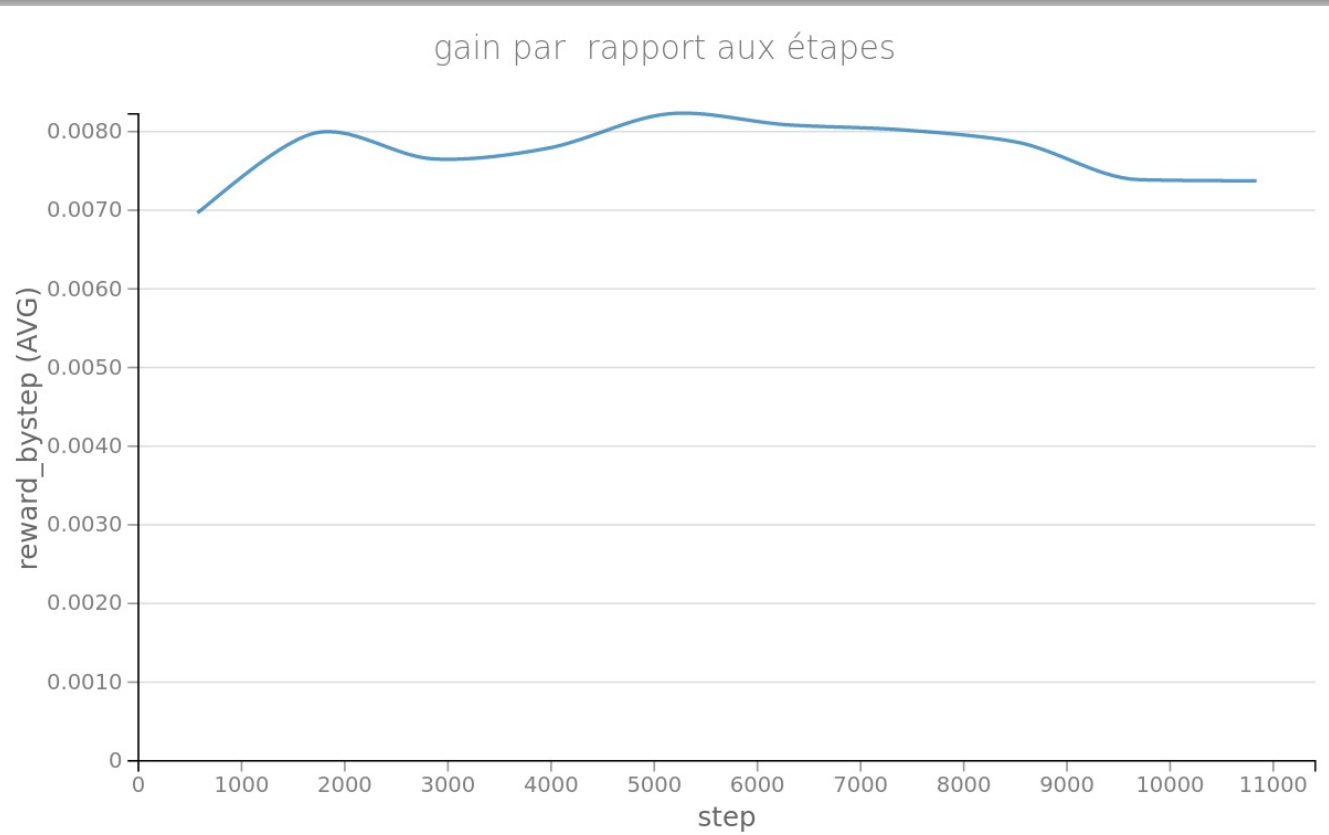

Figure – Courbe du gain en fonction du nombre d'étapes d'apprentissage

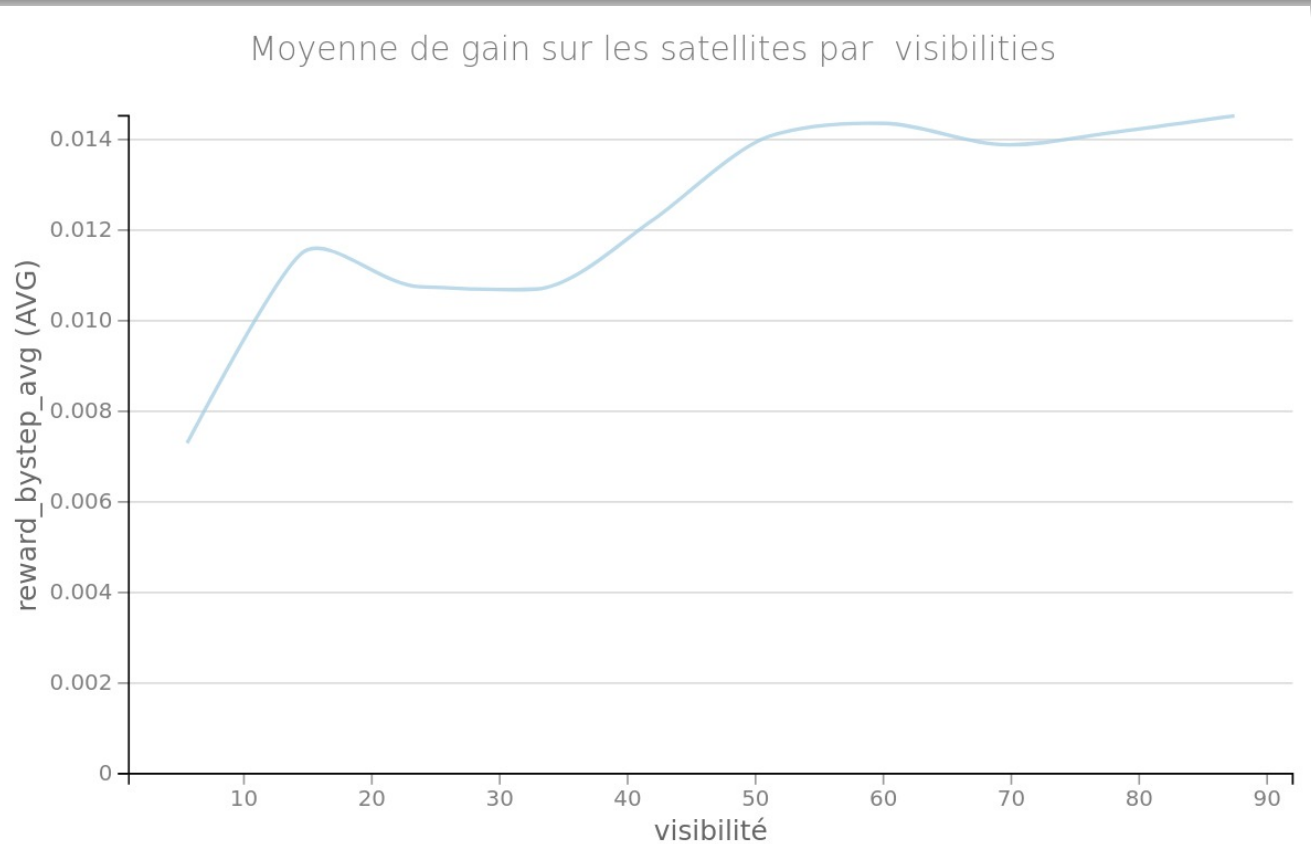

Figure – Courbe de la moyenne du gain sur tout les satellites en fonction du nombre de visibilités station

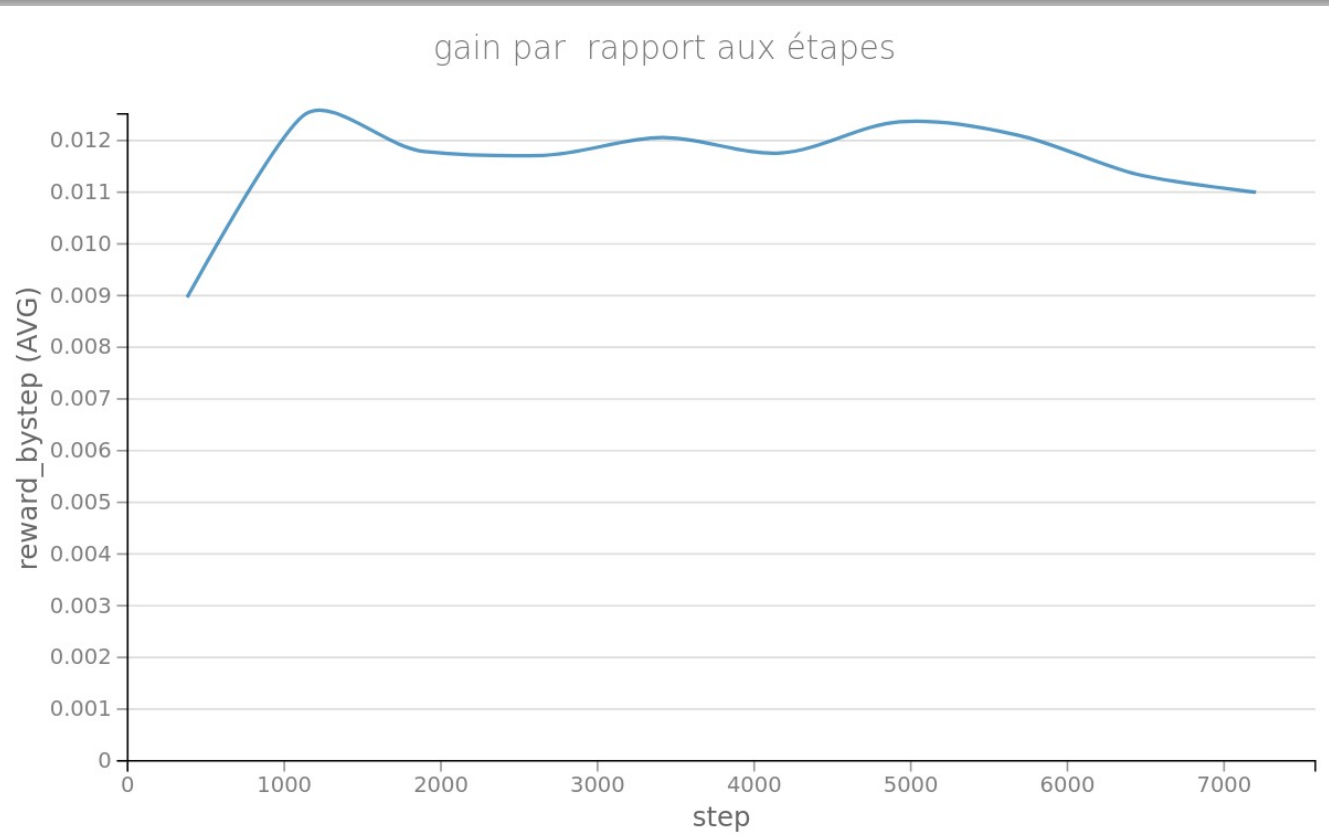

Figure – Courbe du gain en fonction du nombre d'étapes d'apprentissage

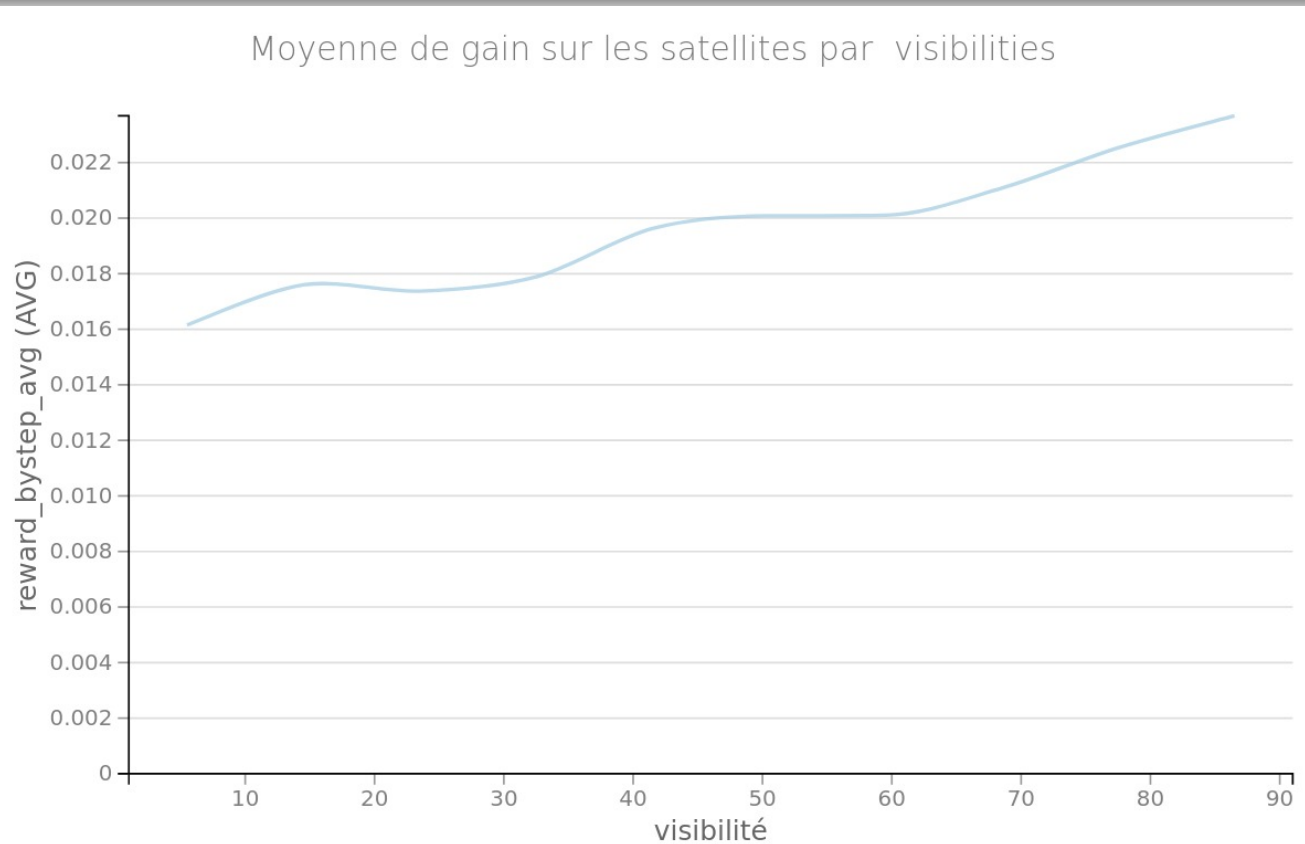

Figure – Courbe de la moyenne du gain sur tout les satellites en fonction du nombre de visibilités station

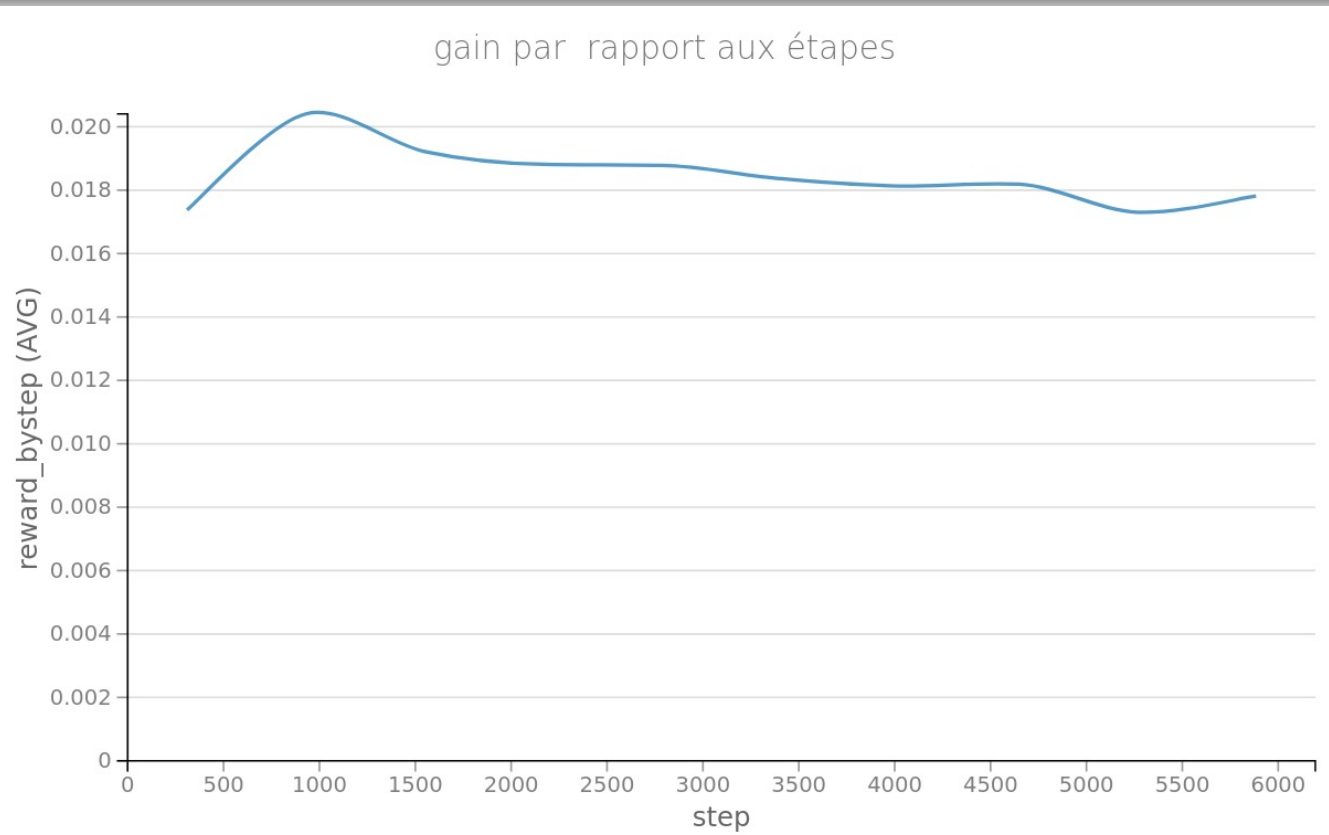

Figure – Courbe du gain en fonction du nombre d'étapes d'apprentissage

Conclusion

Conclusion