Introduction à l'apprentissage par renforcement

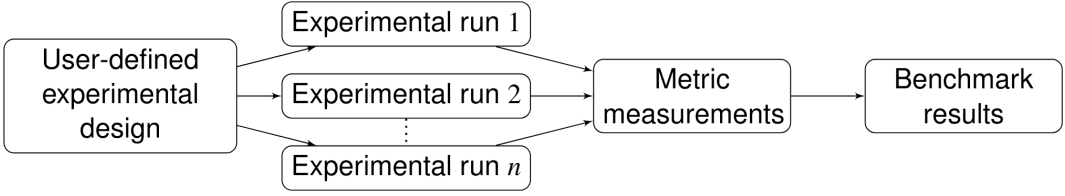
Cours 4 : Benchmarking pour RL

Zhi YAN

ENSTA

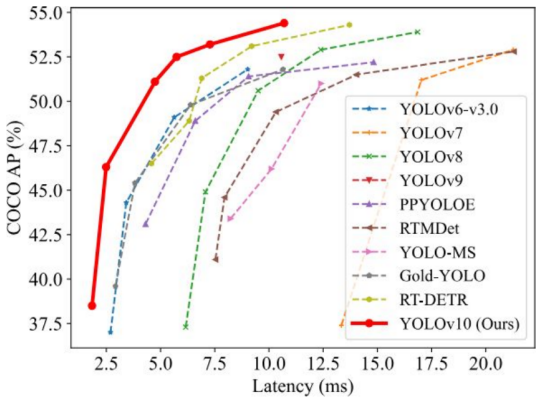
30 janvier 2025

Qu'est-ce que le benchmarking ?



Pourquoi faire du benchmarking ?

- Comparer les performances de différents algorithmes.
- Suivre la progression des algorithmes.
- Découvrir les forces et les faiblesses des algorithmes.
- Guider la sélection des algorithmes.



A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding. YOLOv10: Real-Time End-to-End Object Detection. *NeurIPS*, 2024.

Environnements pour RL

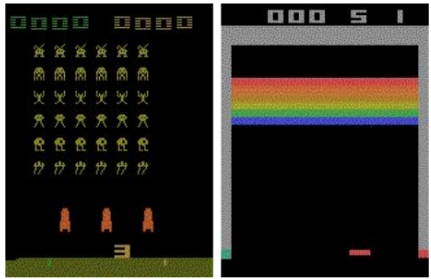
OpenAI Gym :

- Interface standardisée, facile à utiliser.
- Une variété de problèmes de **contrôle classique** : CartPole, MountainCar, Acrobot, Pendulum, etc.
- `gym.make('EnvironmentName')`

Environnements pour RL

Atari Learning Environment (ALE) :

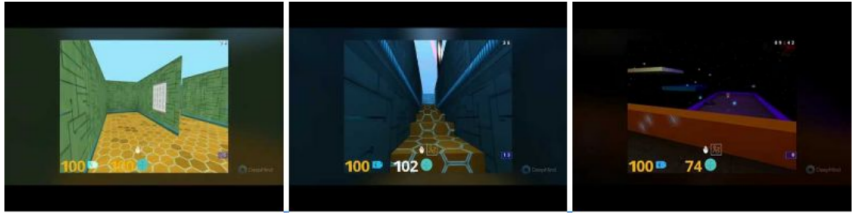
- Une grande collection de jeux Atari 2600.
- Un benchmark important pour les premières méthodes de **RL profond**.
- Entrée visuelle haute dimension.



Environnements pour RL

DeepMind Lab :

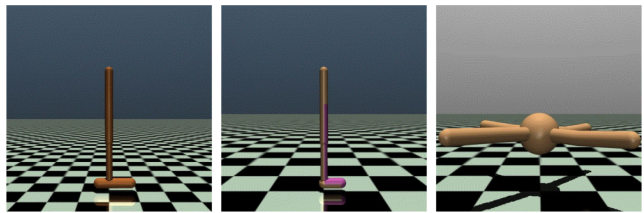
- Construit sur un moteur de jeu 3D.
- Tester les capacités de perception visuelle et de **navigation de l'agent**.
- Plus proche de l'environnement complexe du monde réel.



Environnements pour RL

Autres environnements :

- MuJoCo : Un moteur physique puissant pour les tâches de **contrôle continu**.
- PyBullet : Simulation physique.
- Meta-World : Multi-tâches et méta-apprentissage.



Mesures d'évaluation

Average Return : Mesurer la récompense cumulative moyenne de l'agent sur plusieurs épisodes.

$$AverageReturn = (G_1 + G_2 + G_3 + \ldots + G_m)/M$$

where

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots + \gamma^{(n-t-1)} R_n$$

La mesure d'évaluation la plus couramment utilisée.

ENSTA
IP PARIS

Un exemple

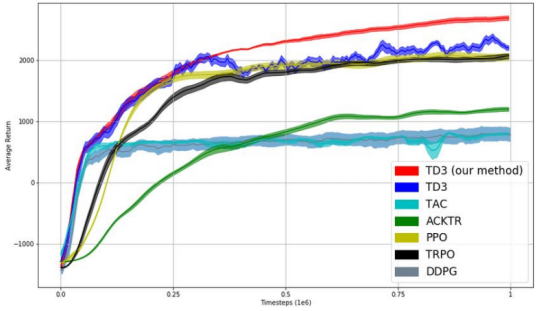
- Supposons qu'un agent ait joué 3 épisodes dans l'environnement CartPole et ait obtenu les récompenses suivantes (avec $\gamma = 0.9$) :
  - Épisode 1 : $R = [1, 1, 1, 1, 1]$ (dure 5 étapes)
  - Épisode 2 : $R = [1, 1, 1, 1, 1, 1, 1, 1]$ (dure 8 étapes)
  - Épisode 3 : $R = [1, 1, 1]$ (dure 3 étapes)
- Alors :
  - Récompense pour l'épisode 1 : $G_1 = 1 + 0.9 + 0.9^2 + 0,9^3 + 0,9^4 \approx 4.095$
  - Récompense pour l'épisode 2 : $G_2 = 1 + 0.9 + 0.9^2 + 0,9^3 + 0,9^4 + 0,9^5 + 0,9^6 + 0,9^7 \approx 5.526$
  - Récompense pour l'épisode 3 : $G_3 = 1 + 0.9 + 0.9^2 \approx 2.71$
- Donc :
  - $AverageReturn = (4.095 + 5.526 + 2.71)/3 \approx 4.11$

ENSTA
IP PARIS

Mesures d'évaluation

Learning Curve :

- Tracer la courbe des récompenses des épisodes à mesure que le temps d'entraînement change.
- Observer le processus d'apprentissage et la vitesse de convergence de l'algorithme.



A. Fayad and M. Ibrahim. Influence-based reinforcement learning for intrinsically-motivated agents. *arXiv preprint*, 2021.

Mesures d'évaluation

- *Success Rate :* Le pourcentage d'épisodes qui accomplissent avec succès une tâche particulière.

$$SuccessRate = \frac{number\ of\ successfully\ completed\ episodes}{total\ number\ of\ episodes}$$

- *Training Time :* Le temps d'entraînement nécessaire à un algorithme pour atteindre une certaine performance.
- *Sample Effciency :* Nombre d'échantillons nécessaires à un algorithme pour atteindre une certaine performance.
- Autres mesures : *Episode Length*, *State Access Frequency*, etc.

Bonnes pratiques pour le benchmarking

- **Choisir le bon environnement de benchmark :** Pertinent en fonction du problème à résoudre.
- **Normaliser les mesures d'évaluation :** Utiliser les mêmes mesures d'évaluation pour comparer différents algorithmes.
- **Contrôler les variables expérimentales :** Contrôler d'autres variables que l'algorithme, telles que les graines aléatoires, les hyperparamètres, etc.

Un exemple

Table 2.2 An example of experimental design for an multi-robot exploration task

Experimental environment	simulator: MORSE [8]
Experimental object	map merging algorithm: probabilistic merging [2]
Experimental dimension	end-to-end and structured
Experimental parameters	robot: Pioneer 3-DX, number of robots: [1, 30] . . .
Parameter vectors	all possible combinations
Experiment repeats	5 times
Experiment stop conditions	99% of the area is explored, the experiment run > 10 mins . . .
Experimental data collection plan	the robot collects data at runtime and stores it locally
Experimental data collection	area explored per simulation step, exploration time when a stop condition is triggered, . . .
Evaluation metrics	exploration time, map quality [27], . . .
Data analysis and results summary	statistically analyze raw data, visualize results using line graphs, . . .

Z. Yan. Robot perception and learning - A human-aware navigation and long-term autonomy perspective. *Springer*, 2025.

Analyse et rapport des résultats

- **Exécuter plusieurs expériences et collecter des statistiques :** Éviter le hasard !
- **Rapporter clairement les résultats expérimentaux :** Utiliser des graphiques et des tableaux.
- **Analyser et discuter les résultats :** Quoi, pourquoi et comment.

ENSTA
IP PARIS

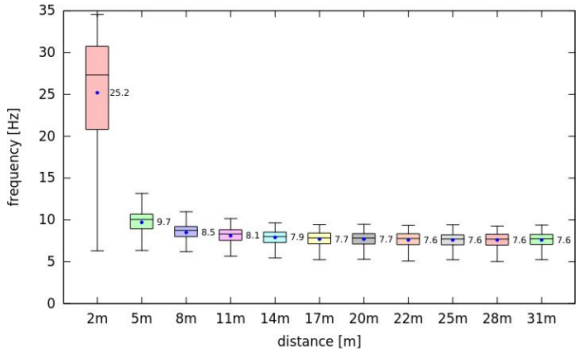
Un exemple

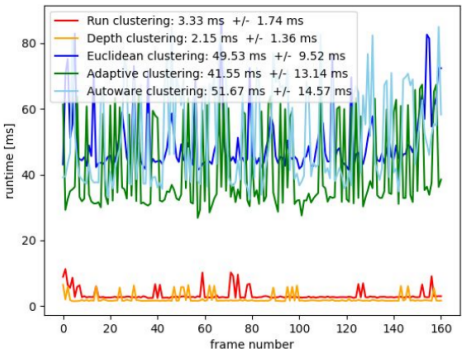
Supposons qu'on comparait les performances de DQN et de REINFORCE dans l'environnement CartPole.

- **Résultats expérimentaux :**
  - DQN a un rendement moyen plus élevé que REINFORCE et converge plus rapidement.
- **Analyse et discussion :**
  - Signification : *Cela montre que DQN obtient de meilleurs résultats que REINFORCE sur cette tâche.*
  - Raison : *L'espace d'état de l'environnement CartPole est petit et la méthode Q-learning utilisée par DQN peut apprendre efficacement la fonction Q optimale. Cependant, la méthode REINFORCE peut rencontrer des problèmes tels qu'une grande variance de gradient sur cette tâche, entraînant une faible efficacité d'apprentissage.*
  - Orientation d'amélioration : *Essayer d'améliorer la méthode du gradient de politique, par exemple en utilisant l'algorithme Acteur-Critique ou en améliorant la méthode d'estimation du gradient.*

ENSTA
IP PARIS

Un autre exemple





Fin