Robotique Social

Homework : Learning from Human Feedback

BOUIN Salomé, CHEN Yuwang
21112382, 21209130

Introduction

Ce projet consiste à effectuer un apprentissage d'un robot dans un environnement Gym à l'aide d'interactions sociaux.
Il existe trois formes d'apprentissage interactif entre l'Homme et le robot : le feedback, la démonstration et l'instruction. Dans ce projet, nous nous intéresserons à l'apprentissage à l'aide de feedback. Ainsi, l'agent agit sur l'environnement en faisant une certaine action, a_t, en fonction de l'état actuelle,s_t et suite à cela, l'homme va évaluer cette action, H(s,a). Ce feedback est considéré comme une observation de la récompense réalisé à t+1 et il s'agit d'une information binaire ou d'un scalaire. Cette évaluation changera alors le comportement du robot afin d'améliorer les performances.
L'exploration de l'environnement faite par le robot est réaliser à l'aide de ces feedback.
Tout d'abord, nous allons effectuer un état de l'art.
Puis, nous allons essayer d'obtenir le résultat espéré d'un des projets trouvés.
Suite à cela, nous allons donner deux types de feedback différents à intégrer dans ce projet.
Ensuite, nous allons comparer nos résultats avec celui de la baseline.
Enfin, nous allons analyser les résultats obtenus.

Problem 1

Benchmark des projets similaires

Table 1: Bachmark

Projet	Algorithme	Environment	Feedback	Score de l'auteur
TAMER	DQL modified	MountainCar	Keyboard	120
Préférence	A2C	Pendulum, Walker, Ant	Preference	-
BCO	IRL	CartPole, MountainCar	Demonstration	-
IRL Minigrid	IRL	Gym-Minigrid	Keyboard	0.9613
HCR Atari	DQL HCR	Montezuma	Human Replay	379.1

Table 2: GitHub

Projet	Lien GitHub
TAMER	https://github.com/benibienz/TAMER
Préférence	https://github.com/mrahtz/learning-from-human-preferences
BCO	https://github.com/montaserFath/BCO
IRL Minigrid	https://github.com/francidellungo/Minigrid$_HCI - project$
HCR Atari	https://github.com/ionelhosu/atari-human-checkpoint-replay

Problem 2

Reproduction des résultat du projet choisi

Dans cette partie, on prend projet TAMER pour les travaux suivants, parce qu'il est le plus facile pour modifier la modalité. Training an Agent Manually via Evaluative Reinforcement (TAMER) va essayer de prédire l'intention de l'Homme, $\hat{H}$(s,a) à l'aide de H(s,a) et d'une méthode de régression. Cette prediction va agir sur la fonction valeur-action de manière suivante: Q'(s,a)=Q(s,a) + β $\hat{H}$(s,a) avec β un facteur de poids décroissant. Cette architecture, comme illustré en Figure1, est basée sur un processus de décision Markovienne sans récompense dont la politique vaut $\pi(s)$= arg $\max_a$ $\hat{H}^*$(s,a).

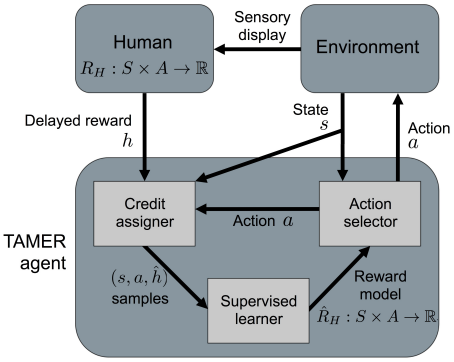

Figure 1: L'architecture du model choisi

D'abord, on refait l'exemple "MountaineCar" dans le projet. L'algorithme Q-learning utilise reward de environnement et TAMER utilise seulement le reward de humain. Mais, on n'obtient pas les même résultats. Une raison est que l'algorithme dans l'article est Sarsa et on a Q-learning dans le projet. Deuxièmement, il utilise les paramètres réglés pour le meilleure reward après 3 et 20 épisodes (Sarsa-3 et Sarsa-20 dans l'article) et les hyperparamètres ne sont pas indiqués dans l'article et dans le projet. Dans l'article, on trouve que Sarsa-20 a une performance similaire que TAMER après 5 époque d'entraînement. Mais, notre methode de Q-learning ne converge pas après 20 époques. Et quand je trouve les autres projet qui utilise SARSA ou Q-learning dans environnement "MountaineCar", il a un reward supérieur à -200 (de manière stable) au moins après 2000 époques. Pour TAMER, notre performance est mille que celle dans l'article. Car TAMER dépend la qualité de feed-back. On peut voir que TAMER converge beaucoup plus vite que Q-learning standard, comme illustré en Figure2.

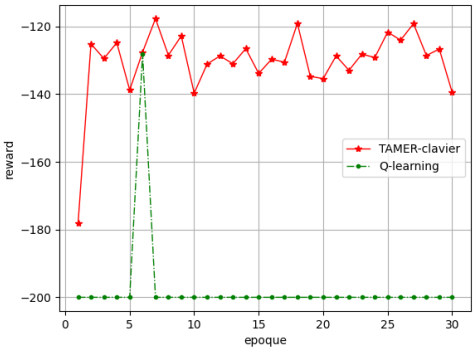

Figure 2: Reward moyen dans MountainCar-V0

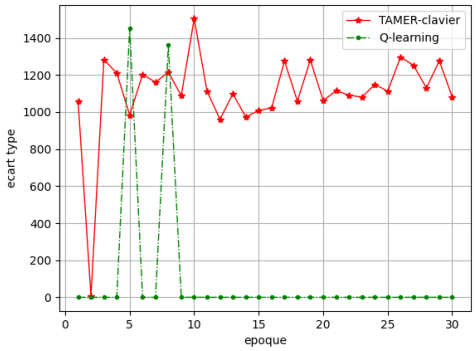

Figure 3: Ecart type de reward

Problem 3

Proposition de deux modalités alternatives pour TAMER

La modalité proposée par le projet original est clavier. Dans cette partie, on propose deux modalités alternatives.

Premièrement, on cherche la modalité de corps (geste/pose). En comparant les méthodes existantes de reconnaissance de geste/pose en temps réel. On trouve que la geste est plus facile à contrôler pour les usagers, son modèle est en Figure4. On choisi MediaPipe.hand comme méthode de détection. Elle fournit les APIs simples pour les développeurs. Comme le reward de humaine est binaire, on cherche la transformation des positions des keypoints des mains à l'indice binaire. On sait la distance entre deux points de la main dépend sa taille dans l'image. Donc, on prend la distance entre point 4 et point 8 comme l_1 et la distance entre point 4 et point 0 comme l_2. Si l_1 est supérieur à l_2, reward est positive. Sinon, il est negative, comme illustré en Figure. Cette méthode est efficace pour acquisition de feedback de humain.

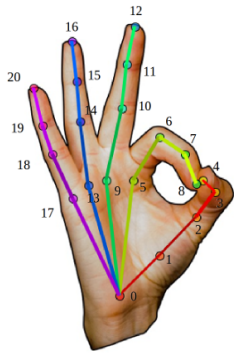

Figure 4: Keypoints de la main

Notre deuxième modalité est parole. Et on a choisi speechrecognization proposé par Google. En

Problem 3 continued on next page. . .

utilisant API de Google, on peut enregistrer l'audio en temps réel et le transforme à text correspond. On prend 'yes' comme un reward positif et 'no' comme un reward négatif.
De plus, on reste dans l'environnement "MountaineCar-V0" avec comme base de référence le Q-learning.

Apprentissage à partir d'instructions

Une instruction est produite par l'humain avec l'intention de communiquer l'action à effectuer dans un état de tâche donné. Cela peut être formulée sous forme d'instruction, qui est une perspective de l'interaction basée sur le langage. Des exemples d'instructions pourraient être : tourner à gauche, ramasser l'objet ou avancer. Ici, on a utilisé clavier au lieu de la parole pour faire les instructions. Avec ce changement, les humains peuvent agir plus vite. On note taper 'r' pour dire à droite, 'l' pour dire à gauche et 's' pour dire arrêt. Dans l'apprentissage avec les instructions original, les instructions vont être transmettes comme la probabilité d'une action. Pour faire la simplification, quand les humains donnent les instructions, l'agent va définitivement prendre les actions correspondantes et mettre à jour les values (Q-learning). Quand il n'y pas d'instruction humaine, il va prendre les actions selon la stratégie $\epsilon - greedy$, comme illustré en Figure5.

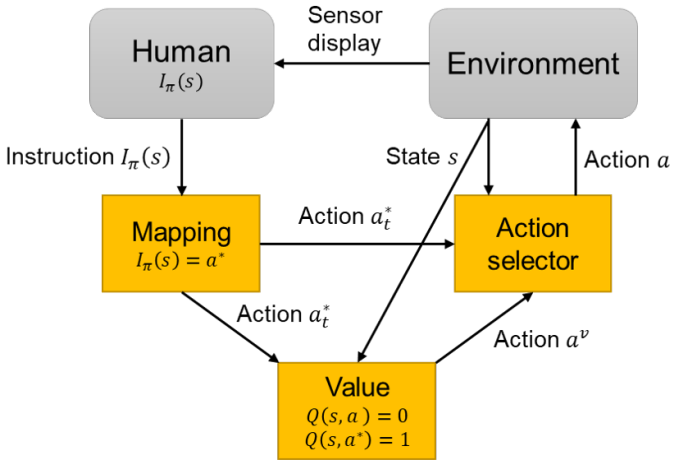

Figure 5: Architecture d'instruction

Cette méthode nous offre les possibilités à l'enseignant par rapport à l'évaluation (feedback) et de montrer (démonstration).

Problem 4

Comparaison avec baseline

Dans cette partie, on fait une comparaison des résultats obtenus avec les approches différentes. La région translucide représente l'intervalle entre le premier quartile et le troisième quartile dans

une époque, comme illustré en Figure6. On trouve que les trois types de feedbacks humains ont tous argumenter la vitesse de convergence comparé avec Q-learning original. En effet, Q-learning original converge après plus de 3000 époques d'entraînement et les approches avec feedbacks humains convergent après 2 ou 5 époques. La performance de l'approche avec la modalité de parole est similaire avec l'approche de geste. On n'a pas un entraînement complet pour cette approche car il prend beaucoup de temps (15 minutes pour un époque avec la présence obligatoire de humain).

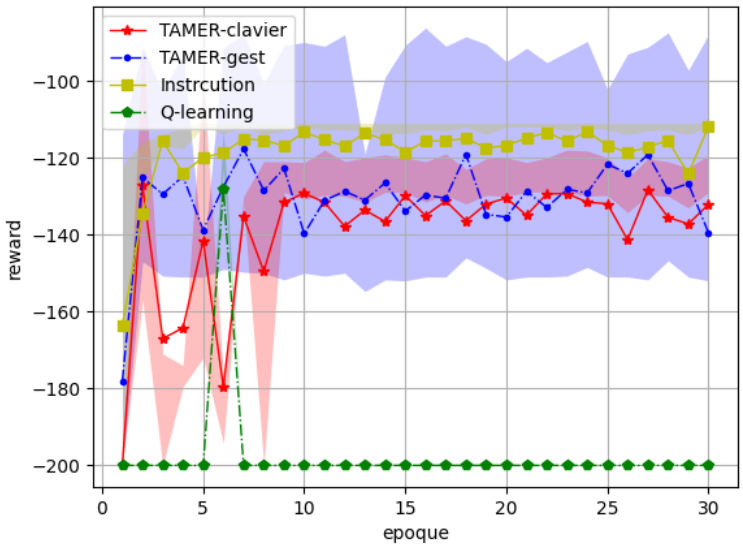

Figure 6: Reward moyen des approches différentes dans MountainCar-V0

Problem 4

Analyse des résultats

Premièrement, si on voit la différence entre la modalité de geste et clavier, les rewards qu'ils a obtenus sont similaires. Mais, la région translucide de la modalité de clavier est beaucoup plus grand que celle de geste, ce qui présente un écart type plus grand, comme illustré en Figure 7. Même on a essayé d'utiliser la même stratégie (politique) pour toutes les approches, notre compétence d'agir varie pour les modalités différentes (fréquence et qualité de feedback). Avec la modalité de reconnaissance de geste, on peut donner les évaluations plus vite. Et cette compétence varie aussi pour les utilisateurs différents. Et puis, on trouve que les moyens de reward de la modalité de reconnaissance de geste est plus petit que le troisième quartile de reward. Car on a des tests qui a un mouvais reward (-200). Même le nombre de tests échoués sont petit comparé avec le nombre total (2/30 ou 3/30), ils ont une influence important pour les moyens des rewards. Les présence des tests échoués sont issue d'incomplétude de l'apprentissage.

Problem 4 continued on next page. . .

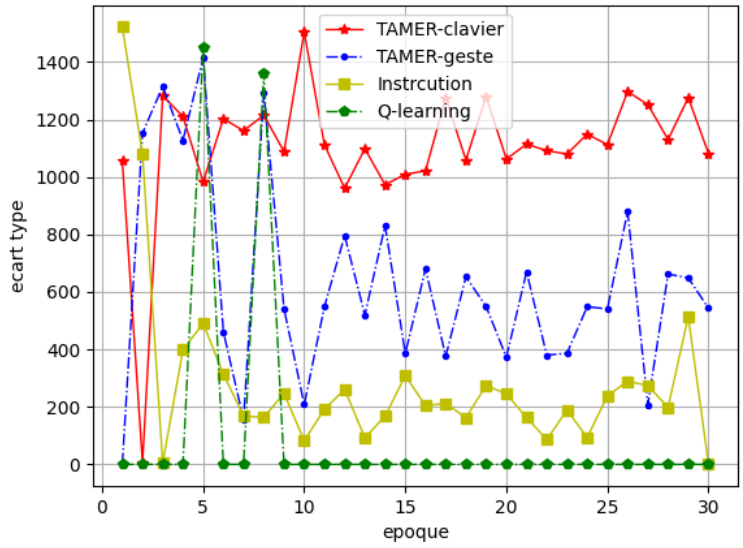

Figure 7: Ecart de Reward des approches différentes dans MountainCar-V0

Comme l'entraînement avec feedback humain est plus court, l'agent ne explore pas tout l'espace état-action. Quand on fait les tests, l'agent peut rentre dans un état inconnu et il a un risque de blocage.
Deuxièmement, la performance de l'approche instruction est meilleure que celle de l'approche TAMER, car on peut présente l'action optimal via l'instruction pour chaque état. L'apprentissage devient plus efficace.

Conclusions

Le feedback humain peut être considéré comme une fonction de récompence, une fonction de valeur ou un politique. Nous avons décidé de le définir comme une fonction de valeur en utilisant l'architecture de TAMER pour la suite de ce projet tout en y ajoutant deux modalités qui sont celle de la gesture et celle de la parole. Après l'application de ces deux modalité. Nous avons développe un méthodes d'instruction simple. Mais, sa performance est la meilleure parmi tous les méthodes que nous utilisons.
Pour l'apprentissage avec feedback humain, la difficulté principale est que les feedbacks humains peuvent être éparses et non-optimaux. Une solution possible est à utiliser l'apprentissage par renforcement interactif basé sur la confiance. La confiance est estimée à partir de l'estimateur de la distribution des valeurs des états et des actions de l'homme et de la distribution des acteurs de l'agent. Nous avons aussi des challenges de ce projet comme déterminer l'engagement, autrement dit de chercher à savoir quand se fini l'intération.

References

[1] Knox, W.B., Stone, P.: Interactively Shaping Agents via Human Reinforcement: The TAMER Framework. In: Proceedings of the Fifth International Conferenceon Knowledge Capture. pp. 9–16. K-CAP '09, ACM, New York, NY, USA (2009).https://doi.org/10.1145/1597735.1597738

[2] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, Dario Amodei: Deep reinforcement learning from human preferences (2017)

[3] Faraz Torabi, Garrett Warnell ,Peter Stone: Behavioral Cloning from Observation (2018)