ECOLE ROYALE MILITAIRE

172ième Promotion POL

Cpt Maxime Sévrin

Année académique 2021 – 2022

2ième Master

La communication entre les agents dans l'apprentissage par renforcement

Par Sous Lieutenant candidat officier de carrière

Yemen RHOUMA

Mémoire de fin d'études du département CISS
présenté en vue de l'obtention du grade académique
de master en sciences de l'ingénieur
sous la direction Capitaine Commandant Koen BOECKX, ir
Bruxelles, 2022

ECOLE ROYALE MILITAIRE
Département CISS

Version 1.0
17/3/2022

La communication entre les agents dans l'apprentissage par renforcement

Yemen RHOUMA

RMA
Royal Military Academy

Avant-propos

Table des matières

Liste des graphiques et figures

Liste des tableaux

1. Introduction

1.1. Contexte

Cette thèse fait partie d'un grand projet "Intelligent Recognition Information System (IRIS)" qui est pris en charge par le département CISS de l'école royale militaire. Le projet a débuté en janvier 2019 et a pour de développer des outils afin d'aider l'équipage des véhicules blindés à exécuter leurs tâches (engager ou non un ennemi potentiel, faire de la reconnaissance ...). Le projet se compose de trois grandes étapes :

- La détection et classification d'objets au sol grâce à des capteurs situés à l'avant du véhicule. Cela conduit à la création d'une grande collection de données sur les objets rencontrés sur le terrain et leurs liens avec le véhicule militaire (position, distance, etc.). à citer SAHARA Semi-Automatic Help for Aerial Region Analysis
- La détection des différentes menaces potentielles afin de créer une carte de ce qui est connu sur le terrain.
- La création d'une stratégie d'attaque à partir de la carte de situation afin de traiter les différentes menaces.

Dans le cadre du troisième point présenté ci-dessus, une partie a été faite par le Cdt Koen BOECKX, ir et qui sera reprise et étudiée en détail. Son travail consiste en :

- Premièrement, le développement d'un modèle algorithmique représentant le terrain. Les acteurs tels que les forces amies et ennemies peuvent interagir avec l'environnement, par exemple en tirant, en se déplaçant, en visant... . L'environnement contient des obstacles qui, par leur présence, peuvent restreindre la visibilité et la mobilité des agents.
- Deuxièmement, voir s'il est possible de créer une stratégie d'attaque basée sur des algorithmes de multi-agents basées sur l'intelligence artificielle et des théorie comme "l'apprentissage approfondi (Deep Learning)" et "l'apprentissage par Renfoncement (Reinforcement Learning)" .

1.2. Analyse de la littérature

1.2.1. L'Apprentissage par Renforcement

L'idée derrière l'apprentissage par renforcement est d'apprendre en interagissant avec l'environnement. Ce type d'apprentissage est celui que les humains expérimentent dès la naissance. Un enfant après la naissance n'a pas d'enseignant explicite. Il dispose de ses différents sens qui lui permettent d'obtenir des informations sur son environnement.

Le domaine de "l'apprentissage automatique" peut être divisé en deux grandes catégories, dont il faut comprendre la différence. Ces deux catégories sont "l'Apprentissage Supervisé (Supervised Learning)" et "l'apprentissage par renforcement".

L'apprentissage supervisé consiste à construire une fonction qui associe une certaine entrée à une certaine sortie sur la base d'exemples. Ces exemples sont étiquetés en fonction de la sortie souhaitée. L'objectif est d'interpréter correctement les nouveaux exemples.

Contrairement au dernier type d'apprentissage, dans l'apprentissage par renforcement, l'agent s'entraîne par l'expérience, non pas en lui donnant une série d'exemples mais en le mettant dans l'environnement. Plus de détails vont suivre dans le paragraphe suivante.

Les éléments de l'apprentissage par renforcement [1]

Il existe deux entités majeures dans l'apprentissage par le renforcement (figure 1.2). Il s'agit de l'**environnement** et de l'**agent**. L'agent est une entité physique ou virtuelle (en simulation) qui est capable d'agir dans un environnement et de le percevoir (de manière limitée). Il est animé par un ensemble de tendances sous forme d'objectifs (fonction d'optimisation) lui permettant de rechercher son but, voire sa survie.

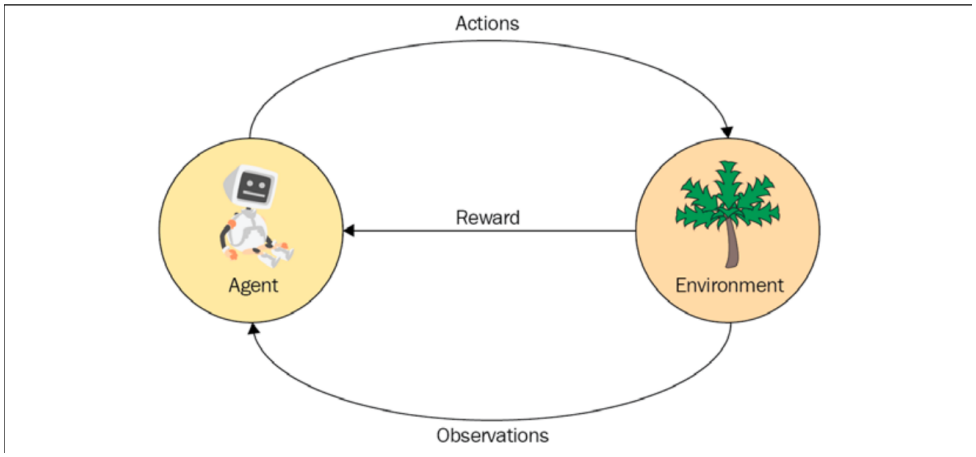

Fig. 1.1. Les entités de l'apprentissage par renforcement [1]

L'agent peut effectuer des **actions** en fonction de l'état actuel qui est défini par l'état de l'environnement et l'état propre de l'agent. Cela conduit à un nouvel état, chaque action entreprise est liée à une récompense ou un coût immédiat (**Reward**). En conclusion, l'objectif de l'agent est d'apprendre quelle action choisir dans chaque état (**observation**) afin d'atteindre son objectif final.

Processus de décision Markovien

Avant d'approfondir la théorie de l'apprentissage par renforcement, il est d'abord nécessaire de définir quelques concepts. Considérons un système dynamique qui est uniquement observable. Les différentes observations forment l'"espace d'état". Pour appeler ce processus un processus de Markov, la propriété de Markov doit être satisfaite. Cette propriété est la suivante : l'observation future d'un système $S(t+1)$ ne dépend que de son état actuel $S(t)$. En bref, seul l'état actuel modélise le système et non toutes les actions précédentes. En mathématiques, un processus de Markov est une séquence de variables aléatoires $(x_0, x_1, x_2..x_n; n >= 0)$ qui satisfont à la condition suivante 1.1 :

$$P(X_{n+1} = x|X_0, X_1, X_2, ..., X_n) = P(X_{n+1}|X_n) \tag{1.1}$$

Cette notion de processus de markov sera étendue par l'introduction de la notion de récompense. La récompense peut être une valeur positive ou négative et peut être grande ou petite. Elle sera ajoutée lors de la transition d'un état à un autre. En outre, la notion d'action est ajoutée au modèle pour modéliser pleinement le problème de l'apprentissage par renforcement.

En conclusion, un processus de décision de Markov est constitué de :

- un ensemble d'états (**States**) , avec un état initial s_0
- un ensemble d'**actions** possibles pour chaque état
- un modèle de transition $P(s'|s, a)$

- une fonction qui détermine la récompense lors du passage d'un état à un autre $R(s)$

La question suivante à se poser est de savoir comment trouver la solution au problème. La solution doit spécifier ce que l'agent est censé faire dans chaque état pour atteindre son objectif. Une telle solution est définie comme la politique de l'agent notée $\pi(s)$.

L'objectif de l'agent est de déterminer la meilleure politique qui conduit à un gain total maximal 1.2 :

$$G_t = R_0 + R_1 + ... + R_n = \sum_{t=0}^{n} R_t \tag{1.2}$$

Dans les processus de décision markoviens qui n'ont pas de but final, la somme des récompenses n'a pas de valeur finie. Pour surmonter ce problème, on introduit le facteur de dévaluation γ. Ce facteur est une valeur comprise entre 0 et 1 afin d'avoir un gain total qui converge. Cela implique que plus les récompenses s'éloignent, moins elles sont importantes. le gain final qui sera maximisé aura l'expression suivante 1.3 :

$$G_t = R_0 + \gamma \cdot R_1 + ... + \gamma^n \cdot R_n = \sum_{t=0}^{n} \gamma^t \cdot R_t \tag{1.3}$$

La valeur de l'état, la valeur de l'action, l'équation de Bellman

La valeur de l'état V(s)
La grande idée derrière l'apprentissage par renforcement est bâtie sur la valeur de l'état V et comment l'approximer à l'aide de l'équation de Bellman. la valeur d'un état s peut être écrite sous cette forme :

$$V(s) = E[\sum_{t=0}^{\infty} r_t * \gamma^t]$$

La valeur de l'action Q(s,a)

La valeur de l'action (Q value) est définie comme la récompense immédiate que l'agent a eu dans un état s plus la récompense qu'il va avoir à long terme jusqu'à avoir atteint son but 1.4.

$$Q(s,a) = E_{s'\ s}[r(s,a) + \gamma \cdot V(s')] \tag{1.4}$$

ceci devient 1.5 :

$$Q(s,a) = r(s,a) + \gamma \cdot max_{a' \in A} Q(s', a') \tag{1.5}$$

Le lien entre V(s) et Q(s,a) est simplet directe 1.6 :

$$V(s) = \max_{a \in A} Q(s,a) \tag{1.6}$$

Les types de l'apprentissage par renforcement

1.2.2. Q Learning

Dans cette section, une technique d'apprentissage par renforcement sera expliquée. Elle permet d'apprendre une stratégie qui indique quelle action entreprendre dans chaque état du système. L'idée est d'apprendre la fonction susmentionnée Q(s,a) qui représente le gain potentiel.
Pour ce faire, on crée un tableau qui détermine le gain pour chaque état du système et pour chaque action possible.

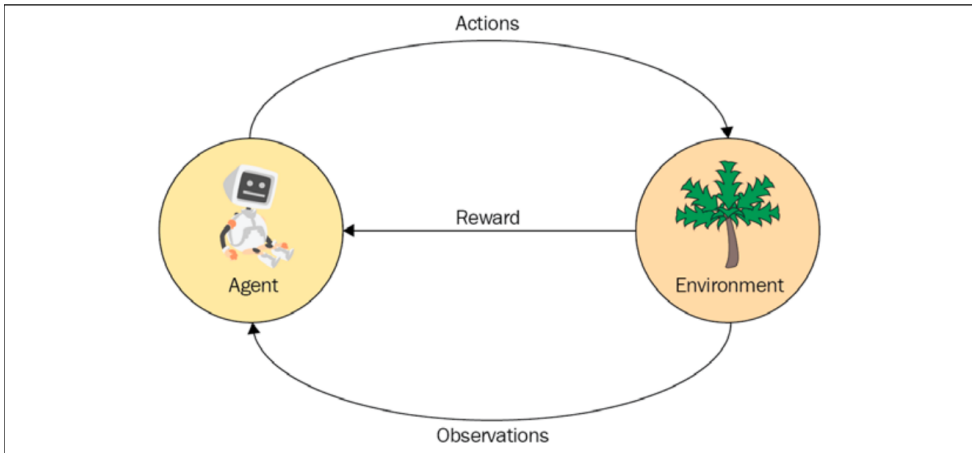

Fig. 1.2. Les entités de l'apprentissage par renforcement [1]

2. Chapter

3. Conclusion

A. Information supplémentaire

Bibliographie

[1] Maxim Lapan. *Deep Reinforcement Learning Hands-On.* Packt Publishing, Birmingham, UK, 2018.