Apprentissage par renforcement profond

Elaboré par Brinssi Alaaeddine
Encadré par Ghazi Bel Mufti

Introduction

En 12 Mars 2016, Le champions du monde et le maître du jeu de Alpha Go Lee Se-dol a perdu un match contre un programme d'intelligence artificielle (IA) en disant j'ai été humilié par la machine.

Grâce à **l'apprentissage par renforcement profond**, la machine a prouvé qu' elle est capable de dépasser l'intelligence humaine , d'où des questions se posent**:**

Problématique

C'est quoi l'apprentissage par renforcement profond?

Comment se distingue l des autres modes d'apprentissage ?

Sur quels principes repose l''apprentissage par renforcement profond ?

Quels sont les domaines d'application auxquels il s'applique ?

Mots-clés

Pour commencer, il faut savoir ces termes là:

Agent : une entité autonome qui agit, orientant son activité vers la réalisation d'objectifs, sur un environnement en utilisant l'observation à travers des capteurs et des actionneurs

Exemples : Robot, Voiture autonome…

État :un état contient les valeurs prises par les variables permettant de localiser l'agent relativement à l'environnement et à ses composants. s_t

Action : les actions correspondent aux comportements possibles que l'agent peut adopter vis-à-vis de l'environnement. .a_t

Exemples : Tourner à gauche, tourner à droite, sauter ...

Récompense : à chaque instant t, l'agent qui se trouve à l'état s_t et choisit d'effectuer une action a_t ,reçoit en contrepartie une récompense r_t qui peut être positive, négative ou nulle. $Q(s_t, a_t)$.

Apprentissage par renforcement : procédure

Le but d'un agent d'apprentissage par renforcement est d'apprendre une politique qui lui permet de gagner le maximum de récompense.

Exploitation/Exploration ou ϵ-greedy

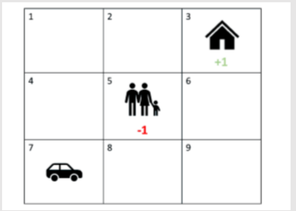
Un agent apprenant est sujet au compromis entre l'exploitation (refaire des actions, dont il sait qu'elles vont lui donner de bonnes récompenses) et l'exploration (essayer de nouvelles actions, pour apprendre de nouvelles choses)

La sélection entre explorer ou exploiter se fait en fonction due la valeur de ϵ qui fait référence à la probabilité d'exploration.

Q-Learning

Le Q-Learning est un algorithme d'apprentissage par renforcement sans modèle permettant d'apprendre la valeur d'une action dans un état particulier..

Pour mieux comprendre, on prend comme exemple un environnement simple de Q-Learning qui est composé d'une grille carrée à 9 cases numérotées de 1 à 9 dont chacune représente un état



Ce tableau présente les récompenses correspondantes à tous les états.

État	Récompense
2	0
3	+1
4	0
5	−1
6	0
7	0
8	0
9	0

En utilisant l'algorithme de Q-Learning, on aura le tableau suivant indiquant la récompense de chaque action dans chaque état

État	$a = 0(\uparrow)$	$a = 1(\downarrow)$	$a = 2(\leftarrow)$	$a = 3(\rightarrow)$
1	0.7501	0.6654	0.727	0.899
2	0.8799	-0.276	0.711	0.999
3	0	0	0	0
4	0.809	0.524	0.621	-0.314
5	0.876	0.116	0.513	0.313
6	0.651	0.025	-0.145	0.166
7	0.725	0.563	0.574	0.221
8	-0.255	0.028	0. 0.468	0.011
9	0.176	0.016	0.001	0.031

Mais, la limite de l'algorithme de Q-Learning c'est que dans la plupart des cas réels , il est difficile de calculer les Q-valeurs. C'est pourquoi on fait appel au réseau de neurone

Deep Q-Learning

Le Q-Learning s'adapte mal aux environnements de grande taille , avec un nombre élevé d'états. D'où, La combinaison de l'idée de Q-Learning et de l'algorithme de réseau de neurones peut également être considérée comme la source d'algorithmes modernes d'apprentissage par renforcement.

L'utilisation de réseaux de neurones avec des problèmes aussi complexes donne les meilleurs estimations de la Q-valeur donnée par l'équation de Bellman . En effet;

Equation de Bellman:

$$Q(s_t, a_t)^{\pi} = r + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1})^{\pi}$$

Les récompenses espérés sachant que a t=0 l'agent est à l'Etat s_t et prend l'action a_t

$$\mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...|s_t, a_t]$$

le réseau de neurone sert donc à modifier les poids de réseau afin que la récompense espérée soit la plus proche possible de celle de l'équation de Bellman ,et au bout de centaines d'épisodes, l'agent sera capable de faire le bon choix en maximisant ses récompenses au sein de l'environnement ,

Outils

La bibliothèque tensorflow
La bibliothèque gym
La bibliothèque random
La bibliothèque numpy
La bibliothèque deque

Conclusion

L'apprentissage par renforcement consiste donc ,pour un agent autonome, à apprendre les actions à prendre, à partir d'expériences, afin d'optimiser une récompense quantitative au cours du temps.

Contrairement aux autres modes d'apprentissage, l'apprentissage par renforcement constitue sa propre données au fur et a mesure que l 'agent apprennent.

Les domaines d'application : les jeux, robotiques, conduite autonome, finance…

Référence

Book: Hands on machine Learning
Thibault Neveu