

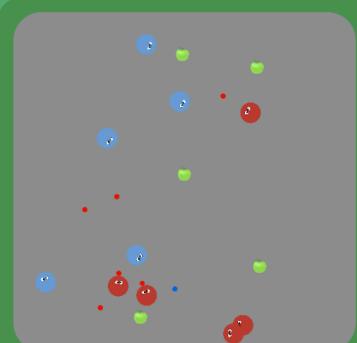
Reinforcement Learning: l'apprentissage machine par planification sur le long terme

L'apprentissage machine nécessite de définir un objectif que la machine doit remplir. Parfois, cet objectif nécessite l'exécution d'une longue suite d'actions, comme les objectifs d'un jeu. Mais comment faire comprendre à la machine que chacune des actions précédentes a eu une influence sur le résultat final ?

Nous avons ici utilisé la méthode de Q-learning, une sous-classe du Reinforcement Learning, qui vise précisément à résoudre ce type de problème. Pour tester notre programme, nous avons créé un environnement de jeu simpliste, où les personnages gagnent des points en mangeant des pommes et en tuant des ennemis, mais en perdent lorsqu'ils sont tués.

Environnement

Dans ce jeu, les agents de chaque équipe gagnent des points en mangeant des pommes ou en attaquant leurs adversaires. Ces agents sont capables d'avancer, tourner ou tirer ; ainsi, "apprendre à jouer" revient à trouver la meilleure action à exécuter dans une situation donnée.



Au départ, les agents ne savent pas quelles actions les mènent à gagner ou perdre des points. C'est en explorant leur environnement qu'ils finissent par découvrir ces informations, puis par développer une stratégie leur permettant de maximiser leur score.

Apprentissage

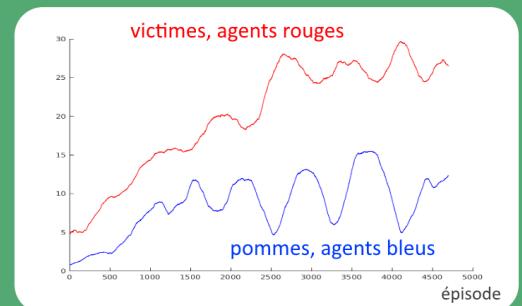
La méthode utilisée ici est connue sous le nom de "Q-Learning" : avant de choisir une action, un agent analyse l'état de l'environnement, défini par certains paramètres : distance et angle du plus proche ennemi, allié, projectile, ou fruit. Cet état correspond à une ligne dans la Q-table, et chaque colonne représente une action possible. Au départ, les cases sont toutes nulles.



L'agent obtient une récompense après chaque action, qu'il ajoute à la case (S, A) correspondante. En plus de cela, il prend la meilleure valeur sur la ligne du nouvel état S' et l'ajoute également à la case précédente. Ceci permet une "propagation" des récompenses : si un agent remplit un objectif plusieurs fois, la récompense de celui-ci va se propager dans les états antérieurs, et il apprendra qu'il peut suivre une certaine suite d'actions pour remplir cet objectif à nouveau.

Résultats

Nous avons adapté nos agents pour qu'ils aient des objectifs distincts : les bleus ne gagnent des récompenses que lorsqu'ils mangent des pommes ; les rouges n'en gagnent que lorsqu'ils tuent les bleus. Les résultats sont concluants : les agents bleus trouvent efficacement les pommes, et les agents rouges ont rapidement appris à les traquer et à les tuer. Voici, au fil des générations, les courbes de pommes mangées par les bleus et d'agents tués par les rouges.



On remarque une claire évolution chez chacune des équipes, bien que les bleus mangent évidemment moins de pommes quand les rouges font davantage de victimes. Néanmoins, ils sont capables de progresser pour survivre dans cette situation, faisant baisser le score des rouges, et ainsi de suite.