Projet IAR – Résumé d'article

L'article « Self-improving reactive agents based on reinforcement learning, planning and teaching. Machine learning » a été écrit par Long-Ji LIN en 1992. Il constate, à son époque, que les algorithmes d'apprentissage par renforcement n'ont été testé que sur des problèmes simples, et que ces derniers convergeaient lentement. L'objet de cette étude était donc d'évaluer les performances de ces algorithmes sur des tâches plus complexes, et de chercher comment il serait possible de rendre cette convergence plus rapide.

Pour cela, il a utilisé deux algorithmes de réseaux de neurones : « adaptative heuristic critic » (AHC) et « Q-learning », ainsi que trois variantes pour ces derniers. La première est « l'experience replay », l'agent garde temporairement en mémoire un historique de ses actions antérieurs, qu'il envoi à son algorithme d'apprentissage plusieurs fois afin d'ajuster les poids des neurones plus rapidement. La deuxième est l'utilisation d'un modèle d'action qui permettrait à l'agent de prédire les conséquences d'une action sans avoir à la faire. La troisième consiste à enseigner manuellement l'agent afin de guider son apprentissage vers la partie plus prometteuse de l'espace de recherche.

Afin de tester ces algorithmes sur une tâche complexe, l'auteur a modélisé un environnement de test composé : d'un agent, d'ennemis, de nourritures et d'obstacles. L'objectif de l'agent est de collecter un maximum de nourritures, tout en évitant les ennemis. L'agent a une vision partielle de l'environnement qu'il observe au travers de différents capteurs, avec des portées différentes, pour la nourriture, les ennemies et les obstacles.

Les tests ont été effectué avec deux représentations des actions possibles de l'agent. Une représentation « globale » dans laquelle les actions consistent simplement à se déplacer vers une case adjacente en utilisant les quatre points cardinaux, et donc sans prendre en compte l'orientation de l'agent. Une représentation « locale » dans lequel on prend en compte cela, les actions possibles sont alors « avancer », « reculer », « se déplacer à gauche », « se déplacer à droite ».

En observant les courbes indiquant le nombre de nourritures obtenues en fonction du nombre de générations, l'auteur remarque que QCON semble être plus performant que AHC avec la représentation globale, mais que les deux algorithmes sont équivalents avec la représentation locale. Il observe également que l'experience replay augmente la vitesse de l'apprentissage, mais que les performances, après un certain nombre de générations (300 ici), sont équivalentes par rapport aux algorithmes de base. Pour l'utilisation d'un modèle d'action, il observe que les algorithmes sont plus performants seulement si le modèle utilisé est suffisamment bon. En ce qui concerne l'enseignement manuelle, il observe que dans le cas global, les performances et la vitesse d'apprentissage sont similaires à la variante experience replay, mais que dans le cas local, cette variante permet d'apprendre encore plus rapidement. Il semblerait que cette variante serait d'autant plus efficace lorsque la difficulté de la tâche augmente.

Les objectifs de notre projet sont de reproduire l'environnement de test décrit, puis de réimplémenter les algorithmes QCON et QCON-R afin d'obtenir des courbes similaires à celles présentées dans l'article.