Carnet de Bord - CEM vs RWR

Réalisé par
Damien LEGROS
Hector KOHLER

Dans le cadre du cours
PANDROIDE

Travail encadré par
Oliver SIGAUD

Master ANDROIDE
Université Sorbonne Université

1 Introduction

Notre sujet concerne les méthodes d'apprentissage qui permettent à un agent (un robot, etc.) d'apprendre une politique c'est-à-dire la manière dont celui-ci doit se comporter dans son environnement pour satisfaire un objectif. Par exemple, les méthodes d'apprentissage permettent à des robots humanoïdes d'apprendre à marcher. Lors d'un cycle d'apprentissage, ces méthodes optimisent une politique initiale jusqu'à ce qu'elle devienne performante. Elles améliorent petit à petit cette politique lors de chaque cycle. En général, les méthodes par renforcement, plus récentes et complexes, performent mieux que les méthodes évolutionnistes, bien plus simples, inspirées du processus d'évolution en biologie. Définissons maintenant les méthodes évolutionnistes. Comme pour l'évolution d'une espèce, petit à petit, des ensembles de politiques sont générées et seules les plus aptes à satisfaire l'objectif sont gardées : c'est la survie du plus apte. Dans le cas de l'apprentissage par renforcement, une représentation mathématique de l'efficacité de la politique (le gradient) est utilisée pour l'optimisation : on modifie ainsi les poids (paramètres du réseau de neurones pour modifier la politique) en fonction de ce gradient. Pour les méthodes évolutionnistes en revanche, on modifie directement les poids et l'on ne passe par par ce gradient. Nous utilisons l'environnement Pendulum : robot moteur devant pousser à gauche ou à droite pour stabiliser une barre en équilibre, sur cette environnement les méthodes évolutionnistes sont les plus performantes. Notre projet consiste donc à expliquer pourquoi les méthodes évolutionnistes sont supérieures aux méthodes d'apprentissage par renforcement dans cette environnement. Pour cela nous avons choisi une approche expérimentale. Pour chaque catégorie nous avons sélectionné une méthode : respectivement CEM et Policy Gradient. Nous comparerons ensuite les performances de ces deux algorithmes sur l'environnement Pendulum et nous utiliserons des outils de visualisation pour regarder les différences.

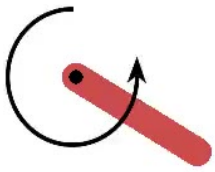

FIGURE 1 – Environnement Pendulum

2 Mots Clés

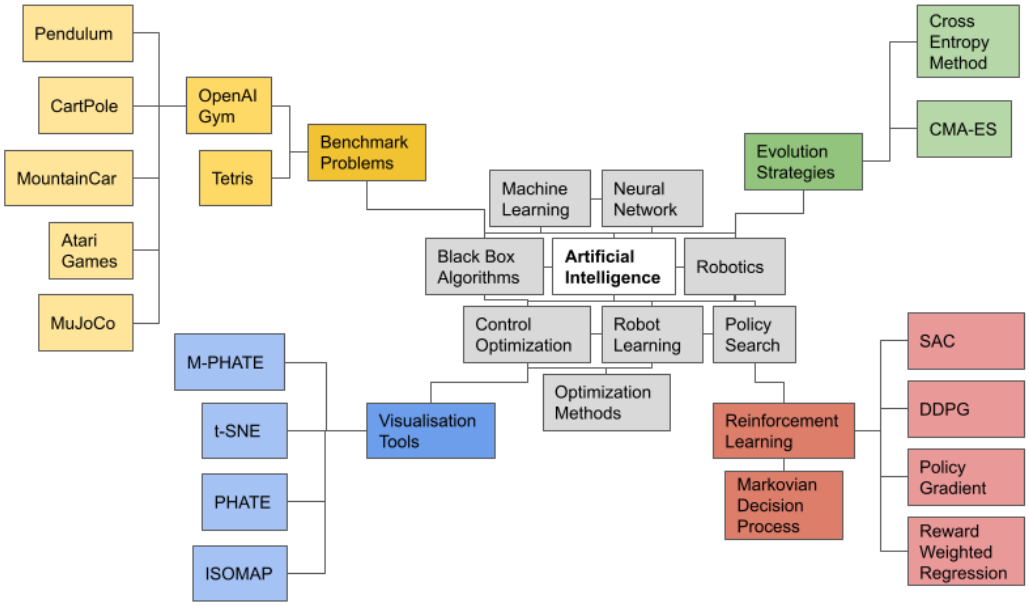

FIGURE 2 – Carte Heuristique non exhaustive des mots clés

3 Descriptif de la recherche documentaire

Pour notre recherche documentaire, nous effectuons tout d'abord une recherche des mots clés sur un moteur de recherche. Cela nous permet d'accéder à des bases de données spécialisées dans l'intelligence artificielle. Les bases de données que nous utilisons sont : arXiv, PMC, ACM Celles-ci permettent d'effectuer une nouvelle recherche des mots clés et d'avoir accès à de nombreux articles. Nous utilisons beaucoup arXiv, car cette base de données est simple d'utilisation et permet d'accéder sans login aux articles. ACM est spécialisée en articles d'informatique, arXiv est spécialisée en articles de sciences et PMC est spécialisée en articles de bio-sciences. Cette dernière base de données est assez éloignée de notre sujet de base. Cependant, les méthodes évolutionnistes sont basées sur des principes de biologie, PMC peut donc être riche en information. Toutes ces bases de données sont spécialisées et fournissent des articles de qualité pour nos recherches. En cas de besoin nous demandons conseil à notre professeur encadrant pour davantage d'informations sur un article ou un sujet. Lorsqu'un article cité nous intéresse, nous effectuons une recherche par rebond en recherchant la référence directement dans Google Scholar. Cela nous permet d'avoir qu'un résultat qui sera l'article en question.

4 Bibliographie

Références

[1] Greg Brockman et al. « OpenAI Gym ». In : *arXiv :1606.01540 [cs]* (5 juin 2016). arXiv : 1606.01540. URL : http://arxiv.org/abs/1606.01540 (visité le 26/03/2021).

[2] Patryk Chrabaszcz, Ilya Loshchilov et Frank Hutter. « Back to Basics : Benchmarking Canonical Evolution Strategies for Playing Atari ». In : *arXiv :1802.08842 [cs]* (24 fév. 2018). arXiv : 1802.08842. URL : http://arxiv.org/abs/1802.08842 (visité le 26/02/2021).

[3] Marc Peter Deisenroth. « A Survey on Policy Search for Robotics ». In : *Foundations and Trends in Robotics* 2.1 (2011), p. 1-142. ISSN : 1935-8253, 1935-8261. DOI : 10.1561/2300000021. URL : http://www.nowpublishers.com/articles/foundations-and-trends-in-robotics/ROB-021 (visité le 26/03/2021).

[4] Yan Duan et al. « Benchmarking Deep Reinforcement Learning for Continuous Control ». In : *International Conference on Machine Learning*. International Conference on Machine Learning. ISSN : 1938-7228. PMLR, 11 juin 2016, p. 1329-1338. URL : http://proceedings.mlr.press/v48/duan16.html (visité le 26/03/2021).

[5] Scott Gigante et al. « Visualizing the PHATE of Neural Networks ». In : *arXiv :1908.02831 [cs, stat]* (7 août 2019). arXiv : 1908.02831. URL : http://arxiv.org/abs/1908.02831 (visité le 26/03/2021).

[6] Shie Mannor, Reuven Y. Rubinstein et Yohai Gat. « The Cross Entropy Method for Fast Policy Search ». In : ICML. 1er jan. 2003. URL : https://openreview.net/forum?id=rkbyRo-_ZH (visité le 26/03/2021).

[7] Paolo Pagliuca, Nicola Milano et Stefano Nolfi. « Efficacy of Modern Neuro-Evolutionary Strategies for Continuous Control Optimization ». In : *Frontiers in Robotics and AI* 7 (28 juil. 2020). ISSN : 2296-9144. DOI : 10.3389/frobt.2020.00098. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7805676/ (visité le 26/02/2021).

[8] Jan Peters et Stefan Schaal. « Reinforcement learning by reward-weighted regression for operational space control ». In : *Proceedings of the 24th international conference on Machine learning*. ICML '07. New York, NY, USA : Association for Computing Machinery, 20 juin 2007, p. 745-750. ISBN : 978-1-59593-793-3. DOI : 10.1145/1273496.1273590. URL : https://doi.org/10.1145/1273496.1273590 (visité le 26/02/2021).

[9] Tim Salimans et al. « Evolution Strategies as a Scalable Alternative to Reinforcement Learning ». In : *arXiv :1703.03864 [cs, stat]* (7 sept. 2017). arXiv : 1703.03864. URL : http://arxiv.org/abs/1703.03864 (visité le 26/02/2021).

[10] Olivier Sigaud et Freek Stulp. « Policy Search in Continuous Action Domains : an Overview ». In : *arXiv :1803.04706 [cs]* (13 juin 2019). arXiv : 1803.04706. URL : http://arxiv.org/abs/1803.04706 (visité le 26/02/2021).

[11] Richard S. Sutton et Andrew G. Barto. *Reinforcement Learning, second edition : An Introduction.* Google-Books-ID : uWV0DwAAQBAJ. MIT Press, 13 nov. 2018. 549 p. ISBN : 978-0-262-35270-3.

[12] István Szita et András Lörincz. « Learning Tetris Using the Noisy Cross-Entropy Method ». In : *Neural Computation* 18.12 (1er déc. 2006), p. 2936-2941. ISSN : 0899-7667. DOI : 10.1162/neco.2006.18.12.2936. URL : https://doi.org/10.1162/neco.2006.18.12.2936 (visité le 26/03/2021).

[13] Tingwu Wang et al. « Benchmarking Model-Based Reinforcement Learning ». In : *arXiv :1907.02057 [cs, stat]* (3 juil. 2019). arXiv : 1907.02057. URL : http://arxiv.org/abs/1907.02057 (visité le 26/03/2021).

[14] Nir Ben Zrihem, Tom Zahavy et Shie Mannor. « Visualizing Dynamics : from t-SNE to SEMI-MDPs ». In : *arXiv :1606.07112 [cs, stat]* (22 juin 2016). arXiv : 1606.07112. URL : http://arxiv.org/abs/1606.07112 (visité le 26/03/2021).

5 Evaluation des sources

— **[7] :** Cet article a été publié en juillet 2020 dans le journal "frontiers in Robotics and AI". L'information est d'un niveau spécialisé, elle nous permet d'estimer les résultats attendus avec les méthodes évolutionnistes suivantes : CMA-ES, xNES, sNES et OpenAI-ES. Les auteurs sont Paolo Pagliuca, Nicola Milano et Stefano Nolfi. Ces personnes sont des chercheurs au "Laboratory of Autonomous Robots and Artificial Life" de "Institute of Cognitive Science" de Rome en Italie. Cet article est diffusé sur NCBI. Les informations sont de qualités et montre les résultats trouvés par ses chercheurs dans différents environnements MuJoCo. Les environnements étudiés sont plus compliqués que celui que nous utilisons. On remarque que les méthodes utilisées sont en général très performantes sur les environnements de MuJoCo.

— **[8] :** Cet article a été publié en juin 2007 lors de la conférence "ICML 2007". L'information est spécialisée et introduit la méthode de Reward Weighted Regression. Les auteurs sont Jan Peters et Stefan Schaal respectivement chercheur à "Institute for Biological Cybernetics" de Tuebingen en Allemagne et chercheur à "University of Southern California" de Los Angeles en Californie. L'information nous permet d'avoir une introduction à la Reward Weighted Regression. Cet article est diffusé sur ACM. L'article est assez difficile à comprendre, mais est cependant intéressant pour visualiser la Reward Weighted Regression. En complément à cet article, notre professeur encadrant nous a fourni un document conçu par lui pour nous expliquer ce que nous devions retenir de RWR et ses similitudes avec Policy Gradient lors de l'apprentissage.

— **[10] :** Cet article a été publié en juin 2019 dans le Journal Neural Networks. L'information est spécialisée et offre un résumé des différentes méthodes de recherche de politiques dans le cas d'un domaine d'action continu. Les auteurs sont Olivier Sigaud, chercheur durant son travail au DLR et Freek Stulp chercheur au DLR, de Wessling en Allemagne. L'information est une mise à jour de 3 précédents surveys datant de 2013 et permet d'avoir un résumé des différentes méthodes tout en ajoutant les nouvelles depuis 2013. Le tableau permet de bien visualiser les méthodes existantes aujourd'hui et leurs relations. Pour plus de détails sur les méthodes d'avant 2013, l'article cite l'article **[3]** que nous avons donc consulté par rebond. Ce document est de plus un survey récent sur les méthodes de Policy Search co-écrit par notre professeur encadrant.