DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **30 novembre 2021**

Nom de famille et prénom de l'auteur : **Monsieur AUBRET Arthur**

Titre de la thèse : *« Apprentissage de compétences de plus en plus complexes via l'apprentissage profond par renforcement en utilisant la motivation intrinsèque »*

Résumé

Introduction

En apprentissage par renforcement (AR), un agent apprend par essais-erreurs à maximiser l'espérance des récompenses reçues suite aux actions effectuées dans son environnement. Ces travaux ont longtemps été limités à des tâches simples jusqu'à ce que de récents travaux intègrent l'apprentissage profond à l'apprentissage par renforcement créant alors le domaine de l'apprentissage profond par renforcement. La capacité de généralisation des réseaux de neurones permet notamment aux algorithmes de résoudre des tâches avec des espaces d'état et d'action à haute dimensionnalité. Par exemple, le Deep Q-Network (DQN) [Mnih et al., 2015] surpasse les performances humaines en utilisant les pixels composant des images brutes.

Malgré ses récents succès, la plupart des travaux en AR profond s'attaquent à des tâches spécifiées par des humains, par exemple, résoudre un jeu atari. C'est une conséquence de la modélisation du problème en AR classique: pour une tâche donnée, un expert exporte la sémantique de la tâche dans une fonction de récompense que l'agent peut maximiser. Typiquement, ca peut être un score lorsque l'agent joue à un jeu atari. Cette fonction de récompense doit être dense et bien structurée de manière à éviter des comportements inattendus [Ng et al., 1999] et donner des informations pertinentes à l'agent.

D'un autre côté, contrairement à l'AR, dès la naissance, les humains apprennent et explorent malgré l'absence d'incitations externes [Ryan and Deci, 2000]. Cette faculté inspira l'émergence du domaine de l'apprentissage développemental [Cangelosi and Schlesinger, 2018, Piaget and Cook, 1952, Oudeyer and Smith, 2016], lequel est basé sur la tendance des bébés, ou plus largement organismes, d'explorer de manière spontanée leur environnement et d'acquérir de nouvelles compétences [Barto, 2013]. Par compétences, nous entendons un comportement qui vise à atteindre un but, qui peut être un ensemble d'états dans l'environnement. En général, les bébés peuvent sélectionner leurs objectifs de manière autonome et les poursuivre en interagissant avec l'environnement [Oudeyer et al., 2013, Baldassarre et al., 2014]. En choisissant des objectifs de plus en plus difficiles à atteindre, un bébé apprend de manière incrémentale et continue de nouvelles compétences et accumule une pléthore de compétences très diverses et de plus en plus complexes. Cette motivation à apprendre est appelée motivation intrinsèque (MI) [Ryan and Deci, 2000].

Dans ce travail, nous faisons l'hypothèse que la motivation intrinsèque peut apporter des éléments clés manquants aux méthodes de l'AR profond. En particulier, nous souhaitons nous émanciper de la supervision d'un expert afin que les agents apprennent de manière autonome et continue à interagir avec leur environnement. Ceci étant fait, les agents pourraient apprendre des comportements de plus en plus complexes sans supervision, comme le font les humains.
L'objectif de ce travail est d'étudier comment tirer profit des motivations intrinsèques pour résoudre les problèmes rencontrés par les algorithmes d'AR profond. Tout au long de notre parcours scientifique, nous donnons d'abord un aperçu des bases de l'AR profond et de la motivation intrinsèque, puis nous détaillons trois contributions qui correspondent respectivement à trois chapitres et nous préservons enfin le dernier chapitre pour une discussion sur les perspectives et les limites de notre travail. Dans ce qui suit, nous résumons les chapitres et nos contributions.

Cadre de l'apprentissage par renforcement et de la motivation intrinsèque

Le premier chapitre est consacré au contexte de l'AR profond et de la motivation intrinsèque. Dans un premier temps, nous passons en revue un algorithme clé de l'AR et son extension à l'AR profond. Dans un deuxième temps, nous nous intéressons à la motivation intrinsèque, nous la définissons d'un point de vue psychologique et nous mettons en évidence

ses propriétés: elle doit être agnostique vis à vis d'une tâche, définie via l'information contenue dans les interactions précédentes d'un agent, dynamique (non stationnaire) et ne doit pas contenir d'information a priori sur l'environnement. Ces propriétés vont nous permettre d'instancier la motivation intrinsèque de manière fructueuse dans le cadre de l'AR. Enfin, nous montrons comment la motivation intrinsèque peut intégrer le cadre de la LR en la reformulant sur la base de l'AR hiérarchique [Sutton et al., 1999], de l'AR conditionné par les objectifs [Schaul et al., 2015] et de la théorie de l'information [Cover and Thomas, 2012]. Cela nous donne l'occasion d'étudier les intérêts de l'utilisation de la motivation intrinsèque en AR.

Etude de la motivation intrinsèque en apprentissage par renforcement

Dans notre première contribution, nous étudions le rôle actuel de la motivation intrinsèque dans le contexte de l'AR. Nous mettons en avant que la motivation intrinsèque permet d'apprendre des compétences abstraites et d'explorer correctement même lorsque les récompenses sont éparses dans l'environnement. Nous analysons et identifions de manière approfondie la façon dont les agents en AR intrinsèquement motivé abordent ces problèmes. Sur la base de cette étude, nous proposons une nouvelle classification de ces méthodes, laquelle est basée sur la théorie de l'information. En particulier, nous formalisons l'apprentissage de compétences, la nouveauté et la surprise et nous en servons pour catégoriser les approches. Par ailleurs, nous proposons d'unifier ces trois types de motivations intrinsèques en formulant l'hypothèse qu'un agent modélise sa cognition de manière hiérarchique; suivant cette hypothèse, la multi-information du modèle permet d'expliquer, seule, une partie du processus d'apprentissage dévelopemental.

Ce travail met en évidence l'importance de l'apprentissage de compétences hiérarchiques avec une motivation intrinsèque. Nous appelons cela une découverte de compétences bottom-up, puisque les compétences sont apprises indépendamment de toute tâche. En bref, l'un de nos résultats est l'observation que l'apprentissage de compétences hiérarchiques bottom-up peut résoudre un grand nombre de problèmes rencontrés par les algorithmes de l'AR profond.

Sur la base de notre analyse, nous supposons qu'un agent peut surmonter les lacunes de l'AR profond s'il apprend séquentiellement des compétences organisées de manière hiérarchique. De cette façon, nous reformulons l'objectif de notre travail : **comment un agent peut-il apprendre des compétences hiérarchiquement organisées de plus en plus complexes avec une motivation intrinsèque ?** C'est la motivation des contributions suivantes.

Apprentissage de bout en bout de compétences réutilisables

Dans notre analyse, nous montrons que les approches précédentes apprennent la plupart du temps les compétences pendant une phase de pré-entrainement non supervisée (ou période de développement), ce qui empêche l'exploration de bout en bout et la spécialisation des compétences pour une tâche ou un objectif donné. S'il n'est pas possible de spécialiser les compétences en fonction d'un but, cela peut également empêcher une intégration dans une structure hiérarchiquement organisée. Sur la base de cette simple observation, nous supposons qu'un agent doit être capable d'apprendre simultanément à résoudre une tâche ou un objectif et à apprendre ses compétences de bas en haut ; il en résulte que son processus d'apprentissage doit être de bout en bout, c'est-à-dire sans phase de pré-entrainement. Pour valider cette hypothèse, nous mettons temporairement de côté l'organisation hiérarchique des compétences et nous nous concentrons sur la manière de découvrir les compétences avec une motivation intrinsèque de bout en bout. C'est la motivation de nos prochaines contributions.

La deuxième contribution est un modèle, End-to-end learning of reusable skills through intrinsic motivation (ELSIM), qui améliore une approche précédente afin d'apprendre un ensemble discret de compétences de bout en bout. En particulier, il construit progressivement un arbre de compétences dans le sens des récompenses extrinsèques.

Cette amélioration ouvre de nouvelles perspectives : 1- l'agent peut maintenant résoudre directement une tâche ou un objectif, ce qui le rend apte à être incorporé dans un cadre hiérarchique ; 2- il peut explorer son environnement même s'il ne reçoit pas de récompenses extrinsèques, ce qui améliore l'exploration. Tout en étant de bout en bout, les compétences peuvent toujours être utilisées dans plusieurs tâches. Nos expériences de preuve de concept valident notre hypothèse.

Cependant, notre analyse approfondie souligne que : premièrement, l'exploration est sous-optimale ; deuxièmement, l'apprentissage d'un ensemble discret de compétences rend difficile son intégration dans une architecture hiérarchique en croissance continue. Ces questions motivent notre troisième contribution.

Découvrir une représentation topologique tout en apprenant des compétences diverses et récompensantes

La troisième contribution est un modèle, Discovering a topological representation to learn diverse and rewarding skills (DisTop), qui apprend une représentation topologique des états-objectifs de son environnement tout en apprenant les compétences pour atteindre ces états-objectifs. Il surmonte les limites d'ELSIM en apprenant des représentations continues et discrètes des objectifs, il peut alors sélectionner les état à visiter selon leur densité (pour l'exploration) ou leur intérêt vis à vis d'une tâche. L'agent apprend toujours de bout en bout, en conservant les meilleures propriétés d'ELSIM.

Dans nos expériences, nous montrons que : 1- DisTop bénéficie des propriétés d'ELSIM, mais il le surpasse clairement sur plusieurs benchmarks ; 2- en oubliant des compétences et en ciblant intelligemment les compétences à améliorer, il obtient des résultats compétitifs avec les méthodes de l'état de l'art sur des tâches à récompense dense, lorsqu'il n'y a pas de tâche et sur des tâches hiérarchiques avec des récompenses extrinsèques rares.

Grâce à notre modèle, nous espérons que d'autres travaux montreront la possibilité et l'intérêt d'utiliser DisTop dans le cadre d'une politique hiérarchique.

Discussion

Pour résumer, nous avons proposé une analyse approfondie des méthodes de motivations intrinsèques en apprentissage par renforcement et avons repéré l'intérêt d'apprendre des compétences avec la motivation intrinsèque tout en apprenant à résoudre des tâches plus haut niveau. Cela nous a mené à introduire deux nouveaux modèles, ELSIM qui apprend des compétences discrètes et DisTop qui apprend des compétences continues. Notre travail présente des perspectives et limites. En particulier, nous pouvons questionner la pertinence de la représentation de DisTop, de nombreuses méthodes d'apprentissage de compétences existent et sont utilisées dans des contextes différents. Le flou sur leurs propriétés et leur difficulté à être démêlées en comparaison des représentations apprises par un VAE [Kingma and Welling, 2013] ouvre la voie à une investigation approfondie. D'autre part, l'unification des motivations intrinsèques sous l'égide de la multi-information semble indiquer que DisTop pourrait bénéficier d'une fonction de cout proche d'autres méthodes. Toutefois, les message qui ressortent de notre travail sont clairs: 1- choisir quelles compétences garder et approfondir rend les modèles d'apprentissages génériques à plusieurs types de tâches; 2- la multi-information d'un modèle cognitif, en unifiant les motivations intrinsèques, peut permettre d'avoir une meilleure appréhension du type de modèle qui pourrait rendre compte d'un apprentissage développemental aboutissant sur la cognition.

References

[Baldassarre et al., 2014] Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M., and Barto, A. (2014). Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. Frontiers in psychology, 5:985.

[Barto, 2013] Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In Intrinsically motivated learning in natural and artificial systems, pages 17–47. Springer.

[Cangelosi and Schlesinger, 2018] Cangelosi, A. and Schlesinger, M. (2018). From babies to robots: the contribution of developmental robotics to developmental psychology. Child Development Perspectives, 12(3):183–188.

[Cover and Thomas, 2012] Cover, T. M. and Thomas, J. A. (2012). Elements of information theory. John Wiley & Sons.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. ArXiv preprint arXiv:1312.6114.4

[Mnih et al., 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540):529.

[Ng et al., 1999] Ng, A. Y., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In ICML, volume 99, pages 278–287.

[Oudeyer et al., 2013] Oudeyer, P., Baranes, A., and Kaplan, F. (2013). Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints. In Baldassarre, G. and Mirolli, M., editors, Intrinsically Motivated Learning in Natural and Artificial Systems, pages 303–365. Springer.

[Oudeyer and Smith, 2016] Oudeyer, P.-Y. and Smith, L. B. (2016). How evolution may work through curiosity-driven developmental process. Topics in Cognitive Science, 8(2):492–502.

[Piaget and Cook, 1952] Piaget, J. and Cook, M. (1952). The origins of intelligence in children, volume 8. International Universities Press New York.

[Ryan and Deci, 2000] Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary educational psychology, 25(1):54–67.

[Schaul et al., 2015] Schaul, T., Horgan, D., Gregor, K., and Silver, D. (2015). Universal value function approximators. In International Conference on Machine Learning, pages 1312–1320.

[Sutton et al., 1999] Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. Artificial intelligence, 112(12):181–211.