Stéphanie Schmidt
14 février 2017

L'intelligence artificielle de Google a appris à devenir "très agressive" lors de situations stressantes.

L'année dernière, le physicien Stephen Hawking avertissait déjà que selon le progrès continu de l'intelligence artificielle (IA), cette dernière sera "la meilleure, ou la pire, des choses qui sera jamais arrivée à l'humanité". Et récemment, l'intelligence artificielle de Google a appris a se comporter de manière "très agressive" dans des situations stressantes.

Les derniers résultats des tests de comportement du nouveau système d'IA DeepMind de Google, montrent clairement que nous devons être prudents lors de la conception des robots de demain.

Lors de précédents tests, l'IA DeepMind a démontré sa capacité à apprendre, indépendamment de sa propre mémoire, et a été capable de battre les meilleurs joueurs de Go au monde.

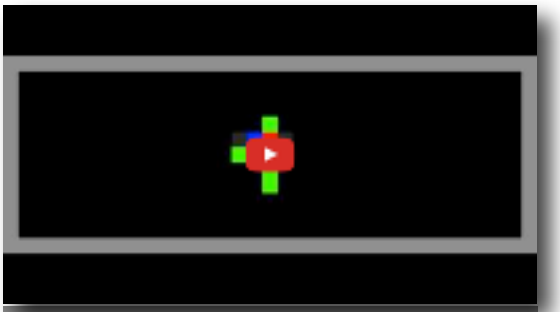
L'IA a également développé la capacité d'imiter sans faille une voix humaine.

Mais à présent, les chercheurs ont testé sa volonté de coopérer avec les autres, et les résultats ont révélé que lorsque l'IA sent qu'elle est sur le point de perdre, elle opte pour des stratégies "très agressives", afin de s'assurer de reprendre le dessus.

L'équipe de Google a effectué 40 millions de parties d'un jeu de "collecte de fruits" simple sur ordinateur, qui demande à deux *agents* de DeepMind de s'affronter, le but étant de rassembler le plus de pommes virtuelles possibles.

L'équipe a constaté que tout se déroulait sans problèmes, tant qu'il y avait bien assez de pommes à récolter. Mais dès que le nombre de pommes a commencé à diminuer, les deux agents sont devenus agressifs, en utilisant des faisceaux laser dans le but de faire sortir du jeu l'adversaire, et récolter toutes les pommes.

Ci-dessous, vous pouvez regarder le jeu de collecte de fruits, avec les deux agents de DeepMind en bleu et en rouge, les pommes virtuelles sont en vert et les faisceaux laser en jaune :



Essayez de regarder cette vidéo sur www.youtube.com

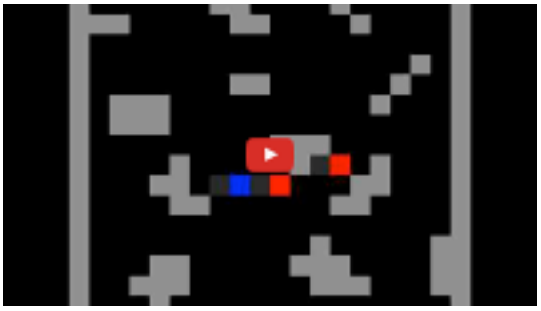
Un fait intéressant est que si un agent arrive à "toucher" son adversaire avec un faisceau laser, aucune récompense supplémentaire n'est donnée. Lorsqu'un agent est touché, celui-ci met un certain temps à revenir dans la partie, donnant l'opportunité à l'autre agent de récolter plus de pommes. Si les deux agents n'utilisaient pas du tout leurs lasers, ils pourraient théoriquement se retrouver avec des parts égales de pommes. C'est l'option qu'ont choisi les itérations "moins intelligentes" de DeepMind.

Ce n'est que lorsque l'équipe de Google a testé des formes de DeepMind de plus en plus complexes, que des éléments de sabotage, de cupidité et d'agression sont apparues. En effet, lorsque les chercheurs ont utilisé des réseaux plus petits comme agents, il y avait une plus grande possibilité de coexistence pacifique. Mais lorsqu'ils utilisaient des réseaux plus grands et plus complexes, l'IA était beaucoup plus disposée à saboter son adversaire, dans le but de le surpasser.

Les chercheurs suggèrent que plus l'agent est intelligent et plus il est capable d'apprendre de son environnement, ce qui lui permet d'utiliser des tactiques très agressives pour prendre le dessus.

> "Ce modèle… montre que certains aspects du comportement humain semblent émerger comme un produit de l'environnement et de l'apprentissage. Des politiques moins agressives émergent de l'apprentissage dans des environnements relativement abondants, avec moins de possibilités d'actions coûteuses. La motivation de la cupidité reflète la tentation de surpasser un rival et de recueillir toutes les pommes soi-même", explique Joel Z Leibo, un membre de l'équipe de recherche.

DeepMind a ensuite été chargé de tester un second jeu vidéo, appelé Wolfpack. Cette fois, il y avait trois agents de l'IA : deux d'entre eux ont eu le rôle des loups tandis que le troisième détenait le rôle de la proie. Contrairement au jeu précédent, celui-ci encourage la coopération : si les deux loups sont près de la proie lorsque celle-ci est capturée, alors les deux reçoivent une récompense, indépendamment de celui qui l'a capturée :



Essayez de regarder cette vidéo sur www.youtube.com

"L'idée est que la proie est dangereuse — un loup solitaire peut la surmonter, mais il risque de perdre la carcasse à cause des charognards. Cependant, lorsque les deux loups capturent la proie ensemble, ils peuvent mieux protéger la carcasse contre les charognards, et donc recevoir une récompense plus élevée", explique l'équipe dans leur rapport.

Concernant le premier jeu, les agents de DeepMind ont appris que l'agressivité et l'égoïsme leur permettait d'obtenir le résultat le plus favorable dans l'environnement en question, mais ils ont également compris que dans le jeu Wolfpack, la coopération était la clé pour un plus grand succès individuel.

Et bien qu'il ne s'agisse que de simples petits jeux informatiques, le message est clair : si vous mettez différents systèmes d'IA en concurrence alors la situation pourrait vite déborder si leurs objectifs ne sont pas équilibrés (dans le but de bénéficier aux humains), car chaque IA souhaiterait alors obtenir le résultat optimal.

"Les systèmes d'IA ont aujourd'hui des capacités impressionnantes mais étroites. Il semble que nous allons continuer à réduire leurs contraintes, et dans le cas extrême, ils atteindront une performance humaine sur pratiquement toutes les tâches intellectuelles. Il est difficile de comprendre à quel point une IA humaine pourrait bénéficier à la société, et il est tout aussi difficile d'imaginer à quel point cela pourrait nuire à la société si celle-ci est conçue ou utilisée de manière incorrecte", a expliqué Elon Musk, fondateur d'OpenAI, une nouvelle association de recherche dédiée à l'éthique de l'intelligence artificielle.

L'équipe de Google n'a pas encore publié son étude car elle doit encore être évaluée par les pairs, mais les premiers résultats (que vous pouvez consulter en fin d'article), révèlent que le simple fait de concevoir des robots et des IA, ne signifie pas qu'ils estimerons nos directives et nos intérêts comme étant une priorité absolue.

D'autres articles sur DeepMind et l'IA :

- ▶ L'intelligence artificielle de Google peut désormais apprendre par elle-même
- ▶ Stopper une intelligence artificielle menaçante ? Google travaille sur un "bouton d'urgence" en cas de problème majeur
- ▶ Source : Multi-agent Reinforcement Learning in Sequential Social Dilemmas