Zeitschrift: Horizons : le magazine suisse de la recherche scientifique
Herausgeber: Fonds National Suisse de la Recherche Scientifique
Band: 31 [i.e. 30] (2018)
Heft: 116

Artikel: Décrypter enfin les rouages de l'intelligence artificielle
Autor: Titz, Sven
DOI: https://doi.org/10.5169/seals-821566

Ceci est un éléphant...

Décrypter enfin les rouages de l'intelligence artificielle

Les prouesses réalisées par les réseaux de neurones impressionnent autant qu'elles intriguent. Des scientifiques tendent des pièges à ces algorithmes afin de tester leurs limites et ainsi tenter de mieux les comprendre. *Par Sven Titz*

Le doute a commencé à se répandre auprès des spécialistes de l'intelligence artificielle (IA). «L'apprentissage automatique est devenu une alchimie», a récemment lancé Ali Rahimi de Google lors d'une conférence récente. La provocation a déclenché un vif débat.

Le contrecoup était probablement inévitable. Les réseaux neuronaux profonds ont enregistré des succès impressionnants au cours des dernières années, en particulier dans la reconnaissance de la parole et des images. Mais ils suscitent désormais un malaise croissant. Capables d'apprendre d'elles-mêmes, ces structures composées de plusieurs couches de neurones artificielles restent opaques. On ne se sait pas vraiment ce qui s'y passe. Peut-on les leurrer? Représentent-elles un risque pour la sécurité? Ces questions ont ouvert des champs de recherche inédits appelés «intelligence artificielle expliquable» ou encore la «neuroscience de l'IA».

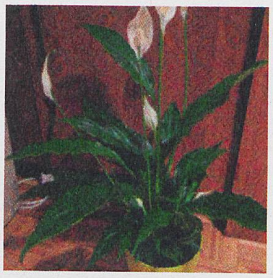
Différents groupes de recherche l'ont montré: il existe de nombreuses possibilités de tromper les réseaux de neurones profonds. Anh Nguyen de l'Université d'Auburn aux Etats-Unis a ainsi produit des images n'ayant pas le moindre sens pour l'homme mais que les algorithmes identifient sans équivoque comme des représentations d'animaux précis.

D'autres méthodes s'avèrent encore plus insidieuses. Des images réalistes sont très légèrement manipulées - l'œil humain n'y voit aucune différence - mais trompent les algorithmes, qui y reconnaissent un objet complètement différent. L'équipe de Pascal Frossard à l'EPFL a par exemple réussi à ce qu'une image de chaussette soit identifiée comme celle d'un éléphant!

«Les systèmes basés sur les réseaux de neurones profonds s'avèrent actuellement plutôt vulnérables aux modifications des données sous-jacentes, note le chercheur. On ne peut souvent pas garantir leur performance.» Un risque pour les voitures autonomes, qui doivent être capables

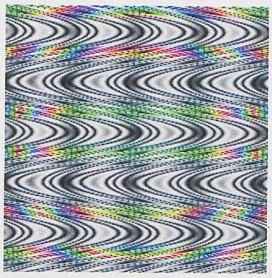
Lorsque les réseaux de neurones se trompent: une chaussette devient un éléphant, une plante un perroquet. Ces images ont été créées par des scientifiques dans le but de tromper les algorithmes de reconnaissance visuelle.

Images: S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard Proceedings of IEEE CVPR, 2017 (élephant, perroquet); Nguyen et al., «DNNs are Easily Fooled», CVPR 2015 (bus, bédé)



... un perroquet

... un bus scolaire



... une bédé.

d'identifier fidèlement les panneaux de signalisation routière sans se laisser tromper par des manipulations, ou encore pour les applications médicales.

Zèbre à huit pattes

Les chercheurs commencent à comprendre d'où proviennent les erreurs. L'une des explications réside dans le nombre limité des exemples utilisés pour l'entraînement des algorithmes. Cela peut mal tourner s'ils se retrouvent confrontés par la suite à des cas entièrement différents. Autre source d'échecs: les réseaux neuronaux n'assimilent pas la représentation structurellement correcte d'un objet. «Prenons un algorithme capable de correctement identifier un zèbre sur une photo, explique Anh Nguyen. Si l'on ajoute à la représentation de l'animal des pattes supplémentaires, l'algorithme sera d'autant plus sûr qu'il s'agit d'un zèbre. Même s'il a huit pattes.»

> «On ne peut souvent pas garantir la performance des réseaux de neurones.»
>
> Pascal Frossard

Le problème: les algorithmes ignorent la structure globale des images, la reconnaissance se basant plutôt sur des détails de forme ou de couleur. Voilà ce que montrent les premières études consacrées à ce qui se passe à l'intérieur des réseaux profonds.

Afin de percer les secrets de l'apprentissage automatique, les scientifiques tels qu'Anh Nguyen recourent notamment à des techniques de visualisation. Ils notent en particulier quels neurones réagissent à différentes propriétés des images. Ces études révèlent qu'en général les premières couches des réseaux de neurones profonds assimilent les caractéristiques de base des exemples utilisés pour l'entraînement, explique Anh Nguyen. Pour les images, il s'agit en particulier des couleurs et des lignes. Plus on pénètre dans le réseau neuronal, plus les informations se combinent. La deuxième couche saisit déjà les contours et les ombres. Ainsi de suite, jusqu'à pouvoir classifier l'objet.

Des parallèles surprenants peuvent être faits avec les neurosciences. Une étude suggère par exemple que des neurones du cerveau se seraient spécialisés dans la reconnaissance de personnalités connues, une hypothèse qui rappelle certains résultats apparus dans l'étude de réseaux de neurones artificiels.

Des concepts théoriques offrent également de nouvelles perspectives telles les propriétés mathématiques des algorithmes, explique Pascal Frossard. Noter si une image est classée comme «pomme» ou comme «poire» permet de délimiter les frontières de décision, à savoir les lignes de démarcation entre diverses catégories d'images.

Mais ces approches, visuelles ou théoriques, éclairent-elles suffisamment le fonctionnement de l'apprentissage automatique? De multiples questions restent ouvertes, répond Yannic Kilcher du Data Analytics Lab à l'ETH Zurich. Elles concernent autant les erreurs que les succès parfois surprenants. Il arrive souvent qu'un programme livre des résultats acceptables même s'il est utilisé pour examiner des données inconnues. «Nous ne comprenons pas encore totalement pourquoi les réseaux neuronaux sont capables de généraliser ainsi», relève-t-il.

Des échecs aux tumeurs

Pour de nombreuses applications, la quantité des données et des paramètres contrôlant les neurones artificiels rend très difficile l'interprétation du comportement des algorithmes. Même les joueurs d'échecs peinent à comprendre les stratégies suivies par les derniers programmes en date, notamment par Google Alpha qui a récemment écrasé le meilleur logiciel d'échecs classique. Mais personne ne sait vraiment comment il a fait. Si des difficultés se présentent déjà pour les échecs, qu'en est-il des programmes médicaux censés aider à classifier les tumeurs? Sont-ils suffisamment compréhensibles et éprouvés pour qu'on puisse se fier à leurs «décisions»? De nombreux chercheurs en doutent - même s'ils n'emploieraient pas le terme «d'alchimie».

La Defense Advanced Research Projects Agency du Département américain de la défense se confronte déjà à ce défi. Son projet Explainable Artificial Intelligence veut développer des réseaux de neurones plus transparents pour les utilisateurs. Des scientifiques de l'Université Stanford ont récemment mis au point un programme qui examine la fiabilité des systèmes d'intelligence artificielle. Il permet de mieux comprendre leurs décisions, réduisant à l'essentiel la complexité de ces modèles.

Pascal Frossard et son équipe poursuivent une autre approche. Ils insufflent des connaissances empiriques préalables dans leurs algorithmes. En combinant l'apprentissage automatique et des connaissances concrètes du réel, ils espèrent pouvoir développer un programme réunissant les avantages des deux approches: les possibilités d'apprentissage des réseaux de neurones profonds et la transparence des programmes classiques. «Finalement, tout dépend des applications, dit le chercheur. Mais le meilleur système se situe probablement quelque part à mi-chemin.»

Sven Titz est journaliste libre à Berlin.

Comment «voler» un algorithme

Une intelligence artificielle peut être dérobée. Non pas physiquement, mais en étant copiée de manière suffisamment fidèle. Les programmes s'entraînent généralement avec des données non publiques, mais une astuce permet néanmoins de reconstituer leur fonctionnement, explique Yannic Kilcher d'ETH Zurich. On leur soumet pour cela des questions (des images dans le cas d'un algorithme de reconnaissance visuelle) qui, combinées avec les résultats, rendent possibles la reproduction du programme à l'aide d'un second réseau neuronal.

Le problème: le réseau neuronal reconstitué est capable de recueillir des informations sur des données d'entraînement privées, un risque particulièrement délicat dans le cas de données confidentielles, par exemple médicales. Des scientifiques, dont Yannic Kilcher, essaient d'introduire des modifications subtiles dans les programmes afin de rendre ces vols de modèles plus difficiles à perpétrer.