mais cela fera globalement progresser la science de l'apprentissage automatique.

Les réseaux de neurones artificiels sont formidables pour une catégorie de tâches particulières qu'Andrew Ng, chercheur principal chez Baidu, entreprise chinoise, décrit comme suit : « Si un individu moyen est capable de réaliser une certaine tâche en moins de une seconde, alors il est probable que vous pourrez l'automatiser. » En revanche, leur mode de fonctionnement et leur programmation par apprentissage à base d'exemples entraînent des caractéristiques dont plusieurs sont des limitations.

ARTIFICIELLE ET… SUPERFICIELLE !

Gary Marcus, professeur émérite de psychologie à l'université de New York, a énuméré dix problèmes liés aux réseaux de neurones profonds, qui nous aident à percevoir ce qu'il faudrait faire pour aller au-delà.

1. L'apprentissage profond exige énormément de données : lors de la phase d'apprentissage, il faut donner au système plus d'exemples que n'en aurait besoin un humain.

2. L'apprentissage à l'aide de réseaux de neurones profonds est en réalité superficiel ; l'utilisation de l'adjectif « profond » concerne la structure des réseaux utilisés, mais pas la nature de ce qui est appris et maîtrisé.

3. L'apprentissage profond ne sait pas, jusqu'à présent, traiter de manière directe les structures hiérarchiques ou impliquant une combinatoire complexe entre éléments (par exemple, le fonctionnement d'un moteur de voiture).

4. L'apprentissage profond est inefficace pour mener des raisonnements logiques enchaînés, comme déduire de « Lucie est la mère de Jacques et la fille de Jean » que « Jacques est le petit-fils de Jean ». On sait construire des méthodes symboliques (non neuronales) qui réussissent ces tâches de raisonnement qu'il faudra probablement associer aux méthodes neuromimétiques.

5. L'apprentissage profond est opaque. Lorsque le réseau a appris à mener une certaine tâche, des milliers de paramètres ont été fixés et, sauf dans les cas les plus simples, il est impossible de les interpréter et de comprendre pourquoi ils ont les valeurs retenues.

6. L'apprentissage profond ne permet pas, jusqu'à présent, d'intégrer des connaissances *a priori* ou formulées abstraitement, par exemple du type : « Si sur le dos de l'animal il y a une selle, il est assez probable que ce soit un cheval. » Les humains ne comprennent pas tout à l'aide d'exemples, ils intègrent aussi des connaissances élaborées par d'autres auxquels ils font confiance, comme leurs enseignants.

7. L'apprentissage profond ne distingue pas une corrélation d'une cause. Il repère des liens, mais n'est pas en mesure de les analyser.

8. L'apprentissage profond suppose un monde stable, ce qui est ennuyeux dans un univers en mouvement ou en évolution rapide.

9. L'apprentissage profond ne garantit pas les résultats. La découverte de méthodes insoupçonnées qui engendrent des exemples pièges en a été une preuve spectaculaire.

10. L'apprentissage profond n'est pas facile à mettre en œuvre, car on ne sait pas à l'avance ce qui marchera, et à chaque nouvelle application, il faut ajuster les paramètres généraux (profondeur du réseau, nombre de neurones, technique d'apprentissage, etc.) sans savoir *a priori* ce qui fonctionnera le mieux.

Vive l'intelligence artificielle, et bravo pour ses succès récents grâce à l'apprentissage profond ! Mais restons lucides sur ce qui est fait et reste à faire avant de parler de machines en tout point meilleures que les humains.

— L'auteur —

> Jean-Paul Delahaye est professeur émérite à l'université de Lille et chercheur au Centre de recherche en informatique, signal et automatique de Lille (Cristal).

— À lire —

> **S. Hussain *et al.*,** Understanding and mitigating audio adversarial examples. 30th USENIX Security Symposium (USENIX Security 21), 2021.

> **L. Fowl, *et al.*,** Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems*,34, 2021.

> N. Akhtar et A. Mian, Threat of adversarial attacks on deep learning in computer vision : A survey, prépublication arXiv:1801.00553, 2018.

> C. Szegedy *et al.*, Intriguing properties of neural networks, *International Conference on Learning Representations*, 2014, prépublication arXiv:1312.6199, 2014.

Entre les performances des réseaux neuronaux et ce qu'on en comprend, c'est le grand écart. Comment la forme d'un réseau conditionne sa fonction est notamment obscur. Les informaticiens travaillent à l'éclairer.

Théorie en voie de construction

Kevin Hartnett

D'un gratte-ciel, nous attendons qu'il respecte un cahier des charges, avec pour *minimum minimorum* de supporter son propre poids et de résister à un séisme. Rien de tel avec l'une des technologies les plus importantes du monde moderne. Avec elle, nous avançons à l'aveuglette, bricolons des configurations, mais tant qu'elle n'a pas été testée, nous ne savons pas vraiment ce qu'elle peut faire, ni où elle échouera. Cette technologie est le réseau de neurones artificiels ; c'est lui qui soutient les systèmes d'intelligence artificielle les plus avancés à ce jour. De plus en plus profondément intégré dans notre société, il filtre le flux des réseaux sociaux que nous recevons, aide les médecins à diagnostiquer des maladies, et influence même la durée d'emprisonnement d'un justiciable condamné pour crime.

En réalité, selon le mathématicien Boris Hanin, « la meilleure approximation de ce que nous savons est que… nous ne savons presque rien à propos du fonctionnement réel des réseaux neuronaux, ni ce que dirait une théorie vraiment pertinente sur le sujet ! » Cet enseignant à l'université de Princeton aime l'analogie avec une autre technologie révolutionnaire : la machine à vapeur. De prime abord, ces machines n'étaient bonnes qu'à pomper de l'eau. Puis, elles ont actionné des locomotives, ce qui correspond peut-être au degré de sophistication atteint aujourd'hui par les réseaux neuronaux. Enfin, les scientifiques et les mathématiciens, grâce à la théorie de la thermodynamique, ont pu comprendre exactement ce qu'il se passe à l'intérieur de n'importe quelle machine. Au final, cette technologie nous a conduits sur la Lune. « D'abord, une belle ingénierie, puis d'excellents trains, puis un peu de compréhension théorique pour en arriver aux fusées », résume Boris Hanin.

Au sein de la communauté tentaculaire des chercheurs en intelligence artificielle, pourtant, un petit groupe d'individus à l'esprit bouillonnant de maths essaye de mettre au point une théorie capable d'expliquer leur fonctionnement et de garantir que, si vous construisez tel réseau neuronal en suivant telle méthode prescrite, il exécutera telles tâches précises. Ce travail n'en est qu'à ses balbutiements, mais, en 2018, plusieurs études ont détaillé la relation entre forme et fonction dans les réseaux neuronaux. L'objectif est d'en définir les fondations mêmes. Il montre que, bien avant de pouvoir certifier qu'ils peuvent conduire des voitures, vous devrez prouver que les réseaux neuronaux sont aptes à réussir une simple multiplication. Voilà qui demande quelques explications.

En bref

> Les réseaux neuronaux ont une architecture inspirée du cerveau sous forme de couches successives de « neurones » reliées entre elles.

> Adapter la profondeur et la largeur des couches est encore très empirique. Il faut beaucoup d'essais et erreurs pour réussir.

> L'un des objectifs majeurs de la discipline est d'établir une théorie sur le lien entre telle forme et tel résultat. Cela nécessite, par exemple, de déterminer la largeur minimale d'une couche.

UN CHIEN, C'EST QUATRE PATTES POILUES

Les réseaux neuronaux visent à imiter le cerveau humain, organe que l'on peut représenter comme une structure qui avance au fil d'abstractions de plus en plus larges. Dès lors, la complexité de la pensée est mesurée par la palette d'abstractions de base sur lesquelles vous pouvez vous appuyer, et le nombre de fois où vous pouvez les combiner en de nouvelles de plus haut niveau – c'est ainsi que nous apprenons à distinguer les chiens des oiseaux.