> Les intelligences artificielles les plus abouties ont aujourd'hui moins de sens commun qu'un rat

Yann LeCun, patron de la recherche en IA de Facebook

Les chercheurs en réseaux neuronaux affirment volontiers que les réseaux « généralisent ». Si un réseau apprend à reconnaître un cheval à partir d'un ensemble de photos de chevaux, il aura la capacité de reconnaître un cheval sur des photos qu'il n'a jamais vues. Il semblait alors aller de soi qu'un réseau qui classe correctement une photo de cheval donnée classera correctement la même photo où quelques pixels seulement ont été modifiés.

L'ÉTRANGE EXPÉRIENCE DE 2014

Ce n'est pas le cas : le réseau indiquera parfois qu'il voit une vache à la place du cheval, que tous les humains identifient pourtant. Si l'hypothèse de continuité « Ce qui est proche d'une image correctement traitée l'est aussi » jusque-là admise était correcte, alors aucune méthode piège de ce type ne pourrait fonctionner. Il est troublant que la technique du groupe de Christian Szegedy fonctionne presque toujours et puisse fabriquer des exemples trompeurs pour une vaste catégorie de réseaux de neurones et de données d'apprentissage.

Pour construire les images pièges, on recourt à une fonction qui estime le risque d'erreur pour chaque image possible, puis on utilise un algorithme d'optimisation qui, de petite modification en petite modification, déforme l'image en augmentant le plus possible ce risque d'erreur, tout en changeant le moins possible l'image manipulée. De proche en proche, sans que l'œil humain ne perçoive de changements notables, l'algorithme construit une image qui fera trébucher le réseau.

Comme souvent en recherche, quand une bonne idée est découverte, une multitude de travaux la précisent et la prolongent. Depuis l'article de 2014, des dizaines d'équipes ont trouvé et perfectionné toutes sortes d'astuces produisant des images pièges, et cela quelle que soit la méthode d'apprentissage mise en œuvre pour instruire le réseau. Dans certains cas, un seul pixel modifié induit en erreur le réseau *(voir l'encadré page 66)*. Assez souvent, une image qui piège un réseau trompera aussi d'autres réseaux ayant appris selon une méthode différente. Christian Szegedy explique : « Nos exemples pièges sont relativement robustes, et trompent aussi des réseaux de neurones ayant un nombre différent de couches que celui qui les a engendrés ou provenant de phases d'apprentissage n'utilisant que des sous-ensembles des données. »

En apprentissage automatique, l'erreur que commettent les systèmes est parfois due à ce qu'on nomme le « surajustement » (*overfitting*). La méthode apprend au système à réagir sur un nombre limité d'exemples, mais dès qu'on en sort, le système se trompe, un peu comme une courbe que l'on force à passer par tous les points de mesure disponibles et qui est incapable de prédire correctement une nouvelle valeur.

Cependant, on a du mal à y voir la bonne explication des exemples pièges, puisque la classification automatique par le réseau fonctionne en général. Ce qui se produit est plus subtil qu'un surajustement : le réseau a généralisé les exemples utilisés dans la phase d'instruction, mais cette généralisation a laissé des failles que détectent les méthodes d'optimisation

employées pour le piéger, en faisant apparaître des cas d'erreur inattendus.

De surcroît, on a su fabriquer des objets 3D qui, quand on les photographie, produisent des images pièges : ce n'est pas une image très spécifique qui trompe le réseau, mais toutes celles qu'engendre un objet réel. C'est inquiétant : on pourrait par exemple fabriquer un faux panneau « Stop » qu'un véhicule autonome interpréterait comme un panneau lui donnant la priorité ! Ces images pièges rendent possibles de nouveaux types d'attaques contre des systèmes informatiques intégrant des réseaux de neurones. Parmi les travaux menés, mentionnons ceux ayant établi qu'on peut piéger plusieurs images à la fois : on cherche à optimiser le pixel qui augmente le plus la mesure du risque d'erreur non pas sur une seule image, mais sur un ensemble d'images. On construit ainsi de proche en proche une perturbation unique, une matrice de valeurs, qui, appliquée à chaque image d'un vaste ensemble d'images, produira une image piège.

Ces découvertes donnent des pistes pour créer de nouvelles méthodes d'apprentissage plus résistantes. L'idée la plus simple consiste, après la première période d'apprentissage, à construire de nombreuses images pièges et à les faire apprendre en complément au réseau en lui indiquant les bonnes réponses, tout comme un professeur de langue indiquerait les faux amis à l'étudiant.

Une autre idée simple fournit facilement une première protection : la compression de données.

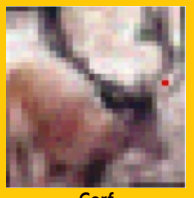
Les images pièges sont conçues pour qu'un réseau ayant appris à les reconnaître propose une réponse incongrue

Gintare Dziugaite, de l'université de Cambridge, et ses collègues ont noté dès 2016 l'intérêt et les effets des méthodes de compression informatique pour se protéger des images pièges. Avant de faire apprendre une image au réseau, on lui appliquera un algorithme de compression d'images d'usage courant. Cela modifie un peu certains pixels, mais la compression, en simplifiant l'image, la débarrasse des caractéristiques inutiles que le réseau pourrait prendre en compte alors qu'elles ne sont qu'accidentelles.

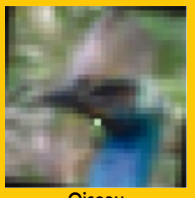
On se retrouve une fois encore dans une situation où le perfectionnement du glaive conduit aux perfectionnements du bouclier. Quelle sera l'issue finale de cette coévolution ? Nul ne le sait,

TROMPER AVEC UN SEUL PIXEL

Les images composées de peu de pixels sont plus délicates à déchiffrer. Pourtant, nous savons le faire assez bien, comme l'illustrent ces six images en basse résolution, assez facilement identifiables (ici un cerf, un oiseau, un chien, un cheval, un bateau, un chat). Dans un de leurs articles, les chercheurs Jiawei Su, Danilo Vasconcellos Vargas et Sakurai Kouichi expliquent comment, après une phase d'apprentissage conduisant à de bonnes identifications, ils ont réussi, en ne modifiant qu'un seul pixel de chaque image (qu'on voit assez nettement), à berner le réseau, qui a alors reconnu autre chose, respectivement un avion, une grenouille, un chat, un chien, un avion et un chien.

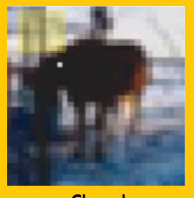


Cerf
Avion

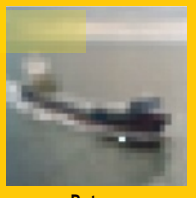


Oiseau
Grenouille

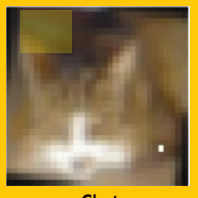
Chien
Chat



Cheval
Chien



Bateau
Avion



Chat
Chien

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.