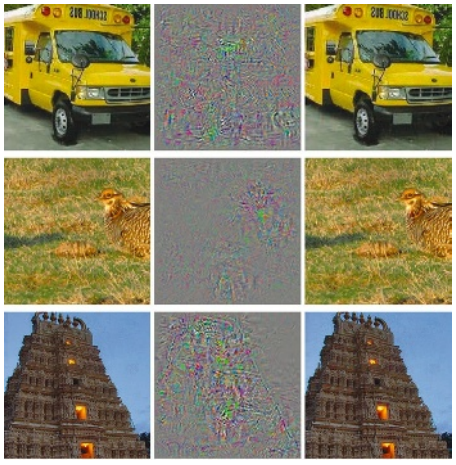
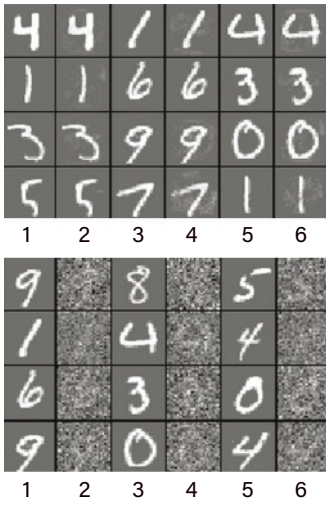
DES IMAGES PIÈGES

Un procédé de modification d'images permet de piéger les algorithmes neuronaux de reconnaissance d'images et les conduit à produire des réponses absurdes.



Les images proposées dans la colonne de gauche de la figure *ci-contre*, après la phase d'apprentissage du réseau de neurones, sont toutes correctement reconnues. Christian Szegedy et son équipe calculent alors des perturbations *(colonne centrale)* qui, appliquées aux images de la colonne de gauche, donnent les images de la colonne de droite. Ces dernières sont alors, dans chaque cas, reconnues par le système comme figurant… des autruches !



Dans le premier tableau de la figure *ci-contre*, les chiffres manuscrits des colonnes 1, 3, et 5 sont identifiés correctement par le réseau de neurones (après apprentissage). En revanche, le réseau identifie mal les mêmes images qui ont subi de légères modifications (colonnes 2, 4 et 6), alors que l'œil humain les identifie aussi bien et ne voit presque pas de différences avec les premières images. Dans le second tableau, on a placé en colonnes 2, 4 et 6 les images des colonnes 1, 3 et 5 soumises à un assez fort bruit aléatoire. Le réseau continue pourtant de fournir les bonnes identifications dans plus de la moitié des cas, alors que les humains en sont incapables.

Ces deux tableaux montrent donc que ce qu'un réseau de neurones apprend dans sa phase d'apprentissage est d'une nature assez différente de ce que nous-mêmes apprenons.

Les chercheurs du domaine étaient persuadés qu'un réseau ayant bien appris était nécessairement robuste et doté d'une propriété de continuité : des traits dessinant de manière vague et approchée une lettre seront reconnus correctement, et une image proche d'une image reconnue sera elle aussi reconnue.

Or ce n'est pas le cas : les chercheurs ont mis au point des images pièges, conçues pour qu'un réseau ayant appris à reconnaître une certaine catégorie d'images propose une réponse incongrue et fausse quand on lui soumet une image proche, qu'un humain classe aisément *(voir l'encadré ci-contre)*.

Insistons sur la prudence nécessaire. Yann LeCun, pionnier de l'« apprentissage profond », qui dirige la recherche en intelligence artificielle chez Facebook, est clair : « C'est un abus de langage que de parler de neurones ! De la même façon qu'on parle d'aile pour un avion comme pour un oiseau, le neurone artificiel est un modèle extrêmement simplifié de la réalité biologique. » Il ajoute : « Les intelligences artificielles les plus abouties ont aujourd'hui moins de sens commun qu'un rat ! »

En somme, les réussites des réseaux de neurones, des méthodes d'apprentissage automatique et des méthodes de fouille de données massives (les *big data*) sont de nouveaux exploits des machines, mais ces succès ne signifient pas que les machines sont devenues véritablement intelligentes.

Revenons à nos exemples pièges. Une équipe de sept chercheurs affiliés aux universités de New York et Montréal, à Google et à Facebook, réunie autour de Christian Szegedy, a publié en 2014 une méthode permettant de tromper les systèmes de reconnaissance visuelle.

> 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