Insistons sur la prudence nécessaire. Yann LeCun, pionnier de l'« apprentissage profond » ayant dirigé 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, mais aussi 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 ! »

Nous disposons de méthodes qui surpassent les humains pour certains types de problèmes et c'est un remarquable succès. Cependant, dans de nombreux domaines, l'informatique a depuis longtemps créé des systèmes qui dépassent les humains. Pouvez-vous mémoriser 300 numéros de téléphone ? Savez-vous calculer 3^{10} en une seconde ? Non. Votre téléphone portable, lui, fait cela facilement. 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.

L'ÉTRANGE EXPÉRIENCE DE 2014

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.

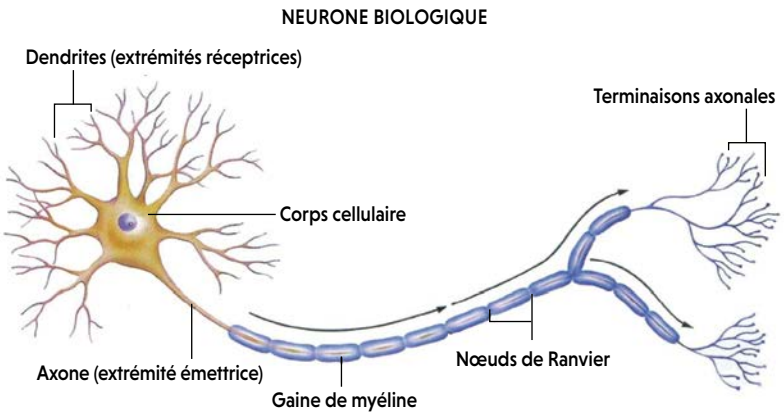
Les chercheurs en réseaux neuronaux affirment volontiers que les réseaux « généralisent ». Si un réseau apprend à reconnaître un cheval à partir d'un ensemble de photos de chevaux, il aura la capacité de reconnaître un cheval sur des photos qu'il n'a jamais vues. Il semblait alors aller de soi qu'un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

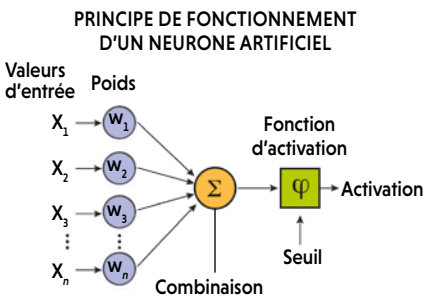
Ce n'est pas le cas : le réseau indiquera parfois qu'il voit une vache à la place du cheval, que tous les humains identifient pourtant. Si l'hypothèse de continuité « Ce qui est proche d'une image correctement traitée l'est aussi » jusque-là admise était correcte, alors aucune méthode piège de ce type 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. Ces chercheurs écrivent : « Pour tous les réseaux que nous avons étudiés et pour toutes les bases d'images utilisées pour l'apprentissage, nous réussissons à engendrer des exemples très proches, visuellement indiscernables d'images correctement reconnues, qui sont mal classées par le réseau. »

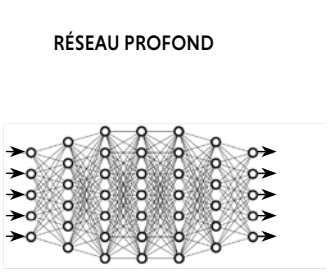
Pour construire les images pièges, on utilise une fonction qui estime le risque d'erreur pour chaque image possible, puis on utilise un algorithme d'optimisation qui, de petites modifications en petites modifications, déforme l'image

2 NEURONES ET RÉSEAUX PROFONDS

Les neurones artificiels sont des versions simplifiées (matérielles ou logicielles) des neurones biologiques. Le plus souvent, on les conçoit de façon que les signaux d'entrée (des influx plus ou moins forts) sont mélangés : on calcule par exemple une combinaison linéaire de ses influx d'entrée en utilisant des paramètres w_1, w_2, ..., w_n propres au neurone. Si cette valeur calculée dépasse un certain seuil (encore un paramètre du neurone), alors un influx de sortie est émis. Un réseau dit profond comporte de nombreuses couches de neurones formels qui interagissent ainsi (schéma du bas, à droite). En réponse aux signaux donnés en entrée (par exemple les pixels d'une image), le réseau produit des signaux en sortie (qui codent par exemple l'objet reconnu dans l'image).







3

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.

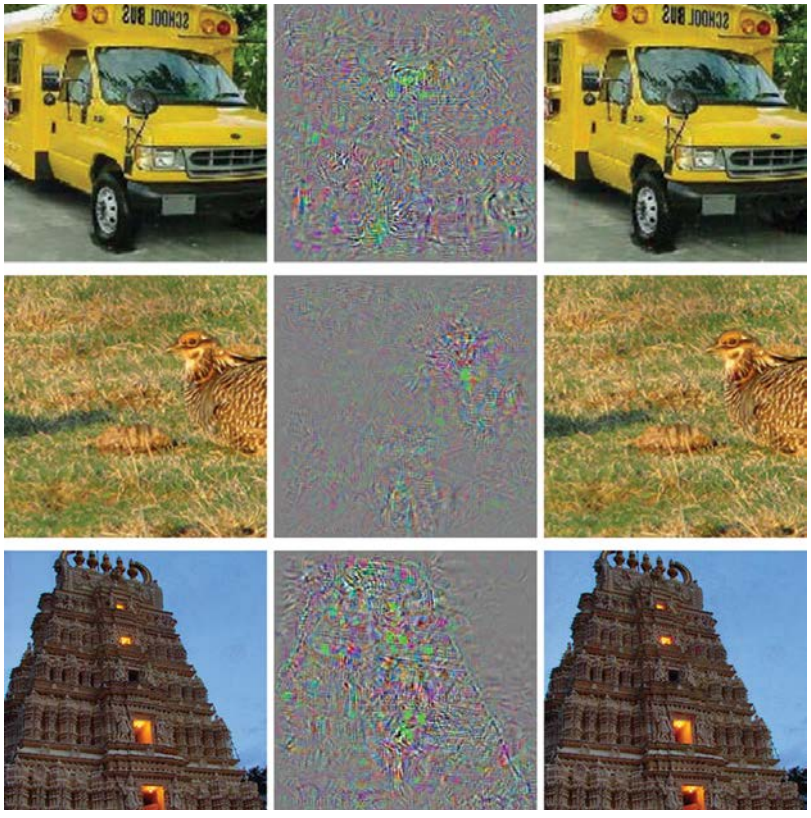
Les images proposées dans la colonne de gauche de la figure *a*, après la phase d'apprentissage du réseau de neurones, sont toutes correctement reconnues. Christian Szegedy et son équipe calculent alors des perturbations *(colonne centrale)* qui, appliquées aux images de la colonne de gauche, donnent 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 *b*, 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.

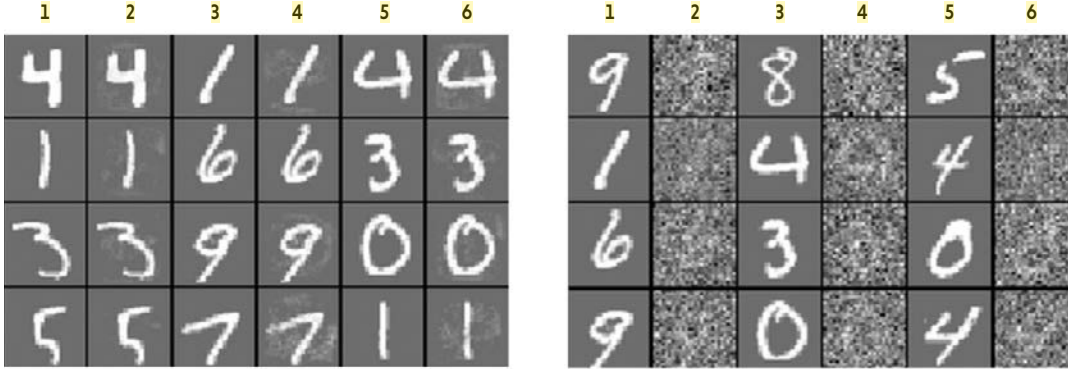
Dans le second tableau de la figure *b*, on a placé en colonnes 2, 4 et 6 les images des colonnes 1, 3 et 5 soumises à un assez fort bruit aléatoire. Le réseau continue pourtant de fournir les bonnes identifications dans plus de la moitié des cas, alors que les humains en sont incapables.

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.

a



b



> 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.

UN NOUVEAU DOMAINE DE RECHERCHE

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 4*). 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.

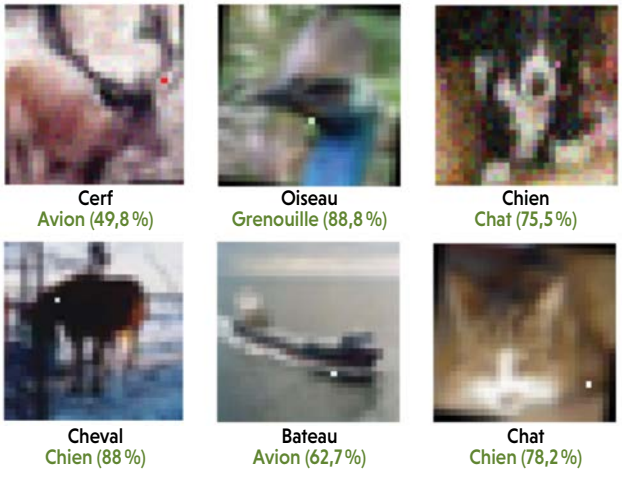
Ce que l'on a réussi à faire pour les images a été adapté aux sons. Les méthodes décrites par Moustafa Alzantot, de l'université de Californie à Los Angeles, et ses collègues permettent de modifier insensiblement le son d'une voix qui prononce «yes» de façon que le système neuronal l'interprétera comme «no» (https://nesl.github.io/adversarial_audio/). On imagine les graves conséquences qu'auraient de telles méthodes si on les utilisait à des fins malveillantes...

Pour évaluer l'effet du bruit ajouté aux sons initiaux pour fabriquer les exemples

4 TROMPER AVEC UN SEUL PIXEL

Les images comportant un petit nombre 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.

Dans l'article « One pixel attack for fooling deep neural networks » (prépublication arXiv:1710.08864, 2017), les chercheurs Jiawei Su, Danilo Vasconcellos Vargas et Sakurai Kouichi expliquent comment, après une phase d'apprentissage conduisant à de bonnes identifications (respectivement un cerf, un oiseau, un chien, un cheval, un bateau, un chat), 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, un chien.



Cerf — Avion (49,8 %)
Oiseau — Grenouille (88,8 %)
Chien — Chat (75,5 %)
Cheval — Chien (88 %)
Bateau — Avion (62,7 %)
Chat — Chien (78,2 %)