L'apprentissage automatique à l'aide de réseaux neuronaux « profonds » est à la mode. Comme souvent quand il s'agit d'intelligence artificielle, un enthousiasme excessif s'est manifesté. Ce fut déjà le cas avec les premiers succès des méthodes d'apprentissage automatique dans la décennie 1950 ; les systèmes experts dans les années 1980 ; puis quelque temps après avec les réseaux de neurones ; et maintenant, c'est une forme de ces réseaux qui fait croire que nous sommes sur le point de mettre dans nos ordinateurs une intelligence générale susceptible de nous surpasser.

Sans dénigrer une remarquable technique, nous aimerions ici remettre sur terre les pieds de ceux qui l'imaginent comme une panacée informatique. Les neurones formels proposés en 1959 sont des modèles informatiques et simplifiés des neurones de notre cerveau. On conçoit des programmes qui en simulent un grand nombre en les regroupant en couches successives, comme les neurones du cerveau humain *(voir les Repères p. 6)*. Les neurones formels de la couche d'entrée du réseau reçoivent des informations, par exemple sous la forme d'images décomposées en pixels : chaque pixel est une entrée. En fonction de leurs paramètres internes, les neurones de cette première couche envoient des signaux aux neurones de la deuxième couche, qui eux-mêmes, en fonction de leurs paramètres internes, envoient des signaux à la troisième couche, etc.

Les influx de sortie, c'est-à-dire de la dernière couche, désignent les réponses possibles. Ce sont par exemple des lettres A, B, C … : si les images données en entrée sont des lettres manuscrites et imprécises, le but recherché est que le réseau, une fois instruit, sache correctement reconnaître les lettres qui lui sont proposées. La phase d'apprentissage consiste à ajuster les paramètres internes de chaque neurone pour que le réseau réponde correctement. Souvent, plusieurs milliers de paramètres sont à calculer. Tout l'art de la programmation des réseaux de neurones consiste, à l'aide d'un grand nombre d'images avec leurs classifications (par exemple des formes manuscrites dont on indique au réseau à quelles lettres elles correspondent), à faire converger les jeux de paramètres des neurones vers un état à peu près stabilisé et qui sera tel que lorsqu'on proposera des images, par exemple d'autres lettres manuscrites non utilisées, elles soient correctement identifiées.

On parle d'apprentissage supervisé : lors d'une phase initiale, on indique au réseau les réponses voulues ; il apprend et, si tout s'est bien passé, devient alors capable de trouver, seul, les bonnes réponses à des entrées nouvelles. Un autre type de méthode pour instruire un réseau de neurones formels se fonde sur l'idée de le récompenser quand il propose une bonne réponse et de le pénaliser dans le cas contraire ; ce sont les méthodes d'« apprentissage par renforcement ».

La puissance de calcul disponible aujourd'hui, les bases de données volumineuses collectées pour les exemples utilisés en phase d'apprentissage, ainsi que la multiplication des couches internes d'un réseau (on en utilise parfois plusieurs dizaines) ont conduit récemment à de spectaculaires réussites, comme la victoire des machines sur les humains au jeu de go.

En bref

> Les succès de l'apprentissage profond laissaient penser qu'un réseau de neurones bien entraîné à reconnaître une catégorie d'images ferait preuve de robustesse.

> Or de légères modifications introduites dans les images parviennent à le tromper, là où l'œil humain n'est pas dupe. La modification d'un seul pixel peut suffire.

> Ces images pièges peuvent exposer des systèmes informatiques à des attaques. Mais elles permettent aussi de forger des parades.

Performances : l'exemple du système CaptionBot

L'identification automatique d'objets sur une photo a fait récemment de grands progrès. En 2018, nous avions testé l'outil CaptionBot, conçu par Microsoft. Le système combine deux réseaux neuronaux. Le premier s'occupe de reconnaissance d'images, le second du traitement du langage naturel pour composer les réponses. En prenant en compte un grand nombre d'images accompagnées de leur description, le logiciel apprend à associer les caractéristiques de l'image à des descriptions écrites de ce qu'elles montrent. Il tente ensuite de faire de même avec les nouvelles images qui lui sont soumises. Voyez ci-dessous les résultats, traduits en français.

BON

Je pense que c'est un arbre avec en fond un coucher de soleil.

BON

Je pense que c'est une personne tenant une guitare.

BON

Je pense que c'est une femme assise sur un canapé.

PAS SI MAL

Je pense que c'est un homme sur une plage.

PAS SI MAL

Je n'en suis pas très sûr, mais je pense que c'est un ours brun assis sur une table.

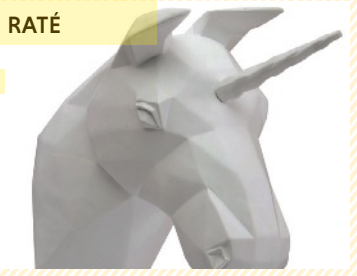
RATÉ

Je pense que c'est un grand étalage de maïs.

RATÉ

Je pense que c'est une paire de chaussures bleues.

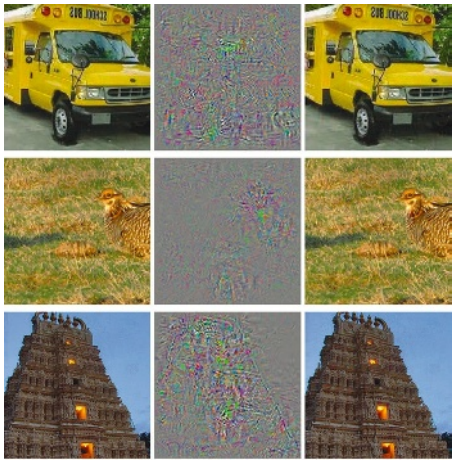
RATÉ



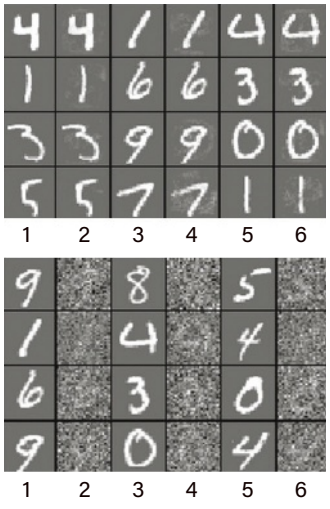
Je n'en suis pas très sûr, mais je pense que c'est un avion dans le ciel.

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.