

PERFORMANCES : L'EXEMPLE DU SYSTÈME CAPTIONBOT

1

L'identification automatique d'objets sur une photo a fait récemment de grands progrès et les systèmes d'aujourd'hui sont capables, avec un certain succès, de reconnaître le contenu d'images délicates. Vous vous en rendez compte vous-même en testant le système en ligne CaptionBot conçu par Microsoft (<https://www.captionbot.ai>).

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 alors de faire de même avec de nouvelles images. Le système continue de progresser. En février 2018, on obtenait par exemple les résultats ci-dessous (les réponses ont été traduites).



BON

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



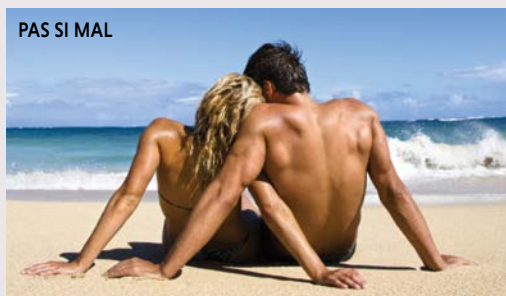
BON

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.

nouvelles. Un autre type de méthodes 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 sinon; 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.

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 fautive quand on lui soumet une image proche, qu'un humain classe aisément.

>

> 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. »

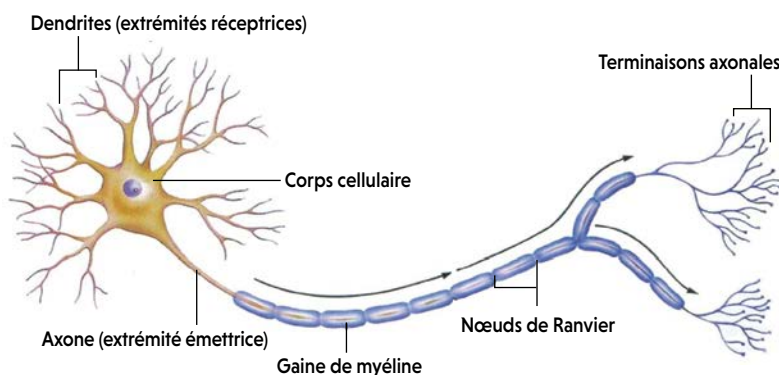
Pour construire les images pièges, on utilise une fonction qui estime le risque d'erreur pour chaque image possible, puis on utilise un algorithme d'optimisation qui, de petites modifications en petites modifications, déforme l'image >

2

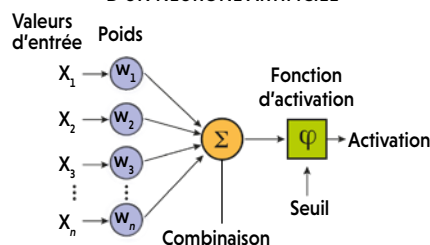
NEURONES ET RÉSEAUX PROFONDS

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 $w_1, w_2, \dots, w_n$ 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).

NEURONE BIOLOGIQUE



PRINCIPE DE FONCTIONNEMENT D'UN NEURONE ARTIFICIEL



RÉSEAU PROFOND

