> 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.

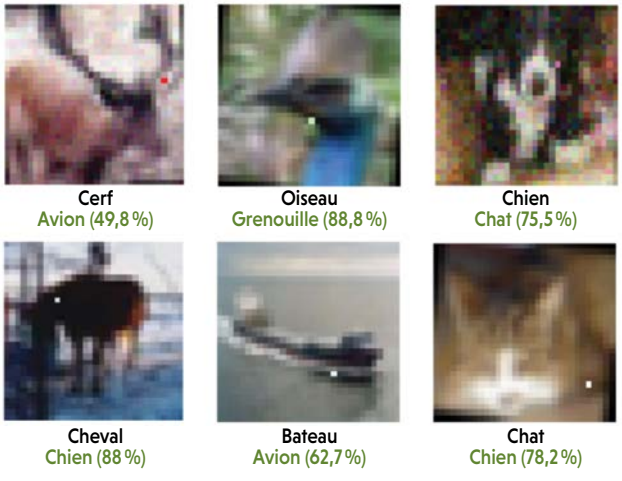
Ce que l'on a réussi à faire pour les images a été adapté aux sons. Les méthodes décrites par Moustafa Alzantot, de l'université de Californie à Los Angeles, et ses collègues permettent de modifier insensiblement le son d'une voix qui prononce «yes» de façon que le système neuronal l'interprétera comme «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 %)