Recherche : quand les réseaux neuronaux perdent les pédales

Des étudiants du MIT (Massachusetts Institute of Technology) ont publié une recherche qui démontre comment des systèmes de vision artificielle peuvent être trompés en identifiant de manière erronée des objets imprimés en 3D.

Ainsi, une tortue imprimée en 3D est l'exemple de ce que les chercheurs appellent une « image contradictoire » (« adversarial image » en anglais). Le classificateur d'images développé par Google appelé Inception-v3 estime qu'elle ressemble plus à une arme à feu qu'à une tortue.

Ces étudiants ont réalisé ces travaux dans le cadre du Labsix, une groupe de recherche en IA du MIT. Cette étude doit être mise dans un contexte où l'IA est de plus en plus sollicitée pour prendre des décisions. Les voitures autonomes en sont la parfaite illustration.

Google et Facebook cherchent des parades

L'équipe d'étudiants du MIT qui a publié la recherche souligne d'ailleurs les risques potentiels liés à ces faiblesses des systèmes de vision artificielle basés sur les réseaux neuronaux : *« Concrètement, cela signifie qu'il est vraisemblable que l'on puisse construire un panneau publicitaire avec des humains qui semble tout à fait ordinaire pour des conducteurs, mais qui pourrait apparaître à une voiture autonome comme un piéton qui apparaît soudainement sur le côté de la rue. Les exemples d'images contradictoires sont une préoccupation pratique que les gens doivent considérer alors que les réseaux neuronaux sont de plus en plus répandus (et dangereux). »* Une méthode baptisée « Expectation Over Transformation » et détaillée dans cet article.

Les éventuels propos alarmistes résultant de la publication de ces travaux doivent toutefois être nuancés. En effet, l'équipe du Labsix indique que la tortue 3D trompe l'IA de Google suivant n'importe quel angle. Or, la vidéo montre qu'il s'agit plutôt de la plupart des angles et non de 100 % d'entre eux.

De plus, le Labsix a eu besoin d'accéder à l'algorithme de vision de Google afin de mettre le doigt sur ses faiblesses et réussir à le tromper. Or, Google et Facebook ont déjà fait savoir qu'ils examinaient les exemples d'images contradictoires du MIT pour trouver la parade et sécuriser leurs systèmes d'IA.