Deep Learning, le grand trou noir de l’intelligence artificielle

Les arbitrages réalisés par les réseaux de neurones sont tellement complexes qu’ils sont impossibles à justifier. Une petite bombe juridique et éthique se prépare.

Temps de lecture : minute

---

20 août 2019

Article initialement publié en novembre 2017

En quelques années, les réseaux de neurones ont supplanté la plupart des autres méthodes d’intelligence artificielle (IA). Plutôt que de chercher à modéliser une vaste quantité d’informations (par exemple calculer tous les coups possibles dans une partie d’échec), ces réseaux de neurones apprennent tous seuls en digérant des millions de données. C’est ainsi qu’AlphaGo, le programme de Google, a battu les meilleurs champions de jeu de Go en 2016 et 2017. Il a observé des dizaines de milliers de parties menées par des joueurs de haut niveau pour devenir lui-même expert.

Cet apprentissage profond (ou Deep Learning) repose sur des couches successives des "neurones" qui effectuent chacun des petits calculs simples. Chaque résultat est transmis à la couche suivante, le niveau de complexité étant de plus en plus élevé. Au final, des résultats bluffants. L’ordinateur est ainsi capable de détecter des tumeurs avant les médecins, prédire les futurs lieux de crimes ou offrir un rendement boursier bien supérieur aux traders humains. Certaines IA sont même en mesure de créer leurs propres musiques originales, d’imaginer un tableau de Van Gogh inédit ou d’inventer un nouveau langage.

" Comme le cerveau, vous ne pouvez pas couper la tête et regarder comment ça fonctionne "

Mais contrairement à un algorithme codé par un chercheur, cette approche rend le fonctionnement du Deep Learning complètement opaque. "Une fois que le réseau de neurones a appris à reconnaître quelque chose, un développeur ne peut pas voir comment il a réussi. C'est comme le cerveau : vous ne pouvez pas couper la tête et regarder comment ça fonctionne", résume Andy Rubin, le cofondateur d'Android, aujourd'hui très impliqué dans l'intelligence artificielle. Et plus le système devient touffu, plus la tâche se corse. "Avec un petit réseau de neurones, il est encore facile de comprendre ce qui se passe", explique Tommi Jaakkola, professeur au MIT. "Mais quand il atteint des milliers de neurones et des centaines de couches, cela devient impossible à décrypter".

Des milliers de textes légaux en déshérence

Et c’est bien là le problème. Car lorsque surgira le moindre incident, personne sera capable de fournir une explication argumentée pour justifier la décision sous-jacente. Comment pourra-t-on se défendre face à un ordinateur vous ayant déclaré "terroriste" sur la base de son propre jugement ? Quid des erreurs de diagnostic pour une maladie grave ? Qui sera jugé responsable en cas d’accident d’une voiture autonome ? La notion de "préjudice prévisible", à laquelle font référence nombre de textes légaux sera impossible à définir dès lors que le comportement d’une IA est, par nature, imprévisible. La Commission européenne accuse par exemple Google de favoriser les résultats de son propre comparateur de shopping dans ses résultats de recherche. Une charge qui sera demain impossible à démontrer si c’est un algorithme auto-apprenant et non un humain qui alimente le moteur de recherche.

Instaurer une " responsabilité algorithmique "

Face à ce casse-tête juridique et éthique à venir, la mobilisation grandit dans les rangs des entreprises et des autorités. La DARPA, l’agence de recherche militaire américaine, a alloué 6,5 millions de dollars en mai 2017 à des chercheurs pour permettre à l’IA de fournir une "explication visuelle ou écrite" de ses décisions. En 2016, le patron de Microsoft Satya Nadella a lui appelé à une "responsabilité algorithmique". Le futur règlement européen sur les protection des données, censé entrer en vigueur en 2018, prévoit que les entreprises utilisant "toute forme de traitement automatisé de données [...] servant à évaluer ou prédire certains aspects de la vie personnelle" puisse fournir des informations sur "la logique sous-jacente" de l’algorithme.

Tenter de percer la " boite noire "

Sauf que pour l’instant, même les meilleurs informaticiens sont incapables d’expliquer quoi que ce soit. Pour tenter de percer les mystères de ces réseaux géants, des chercheurs de Google on "inversé" le processus de reconnaissance d’image, en demandant à l’algorithme de générer ses propres représentations. Résultat : des "animaux nuages" aux allures d’hallucinations, des chiens apparaissant au milieu de arbres... "Si un nuage ressemble un peu à un oiseau, alors le système va le faire ressembler encore plus à un oiseau", expliquent les ingénieurs. "En réitérant l’action, le programme va reconnaître un oiseau plus fortement et ainsi de suite jusqu’à ce qu’un oiseau très détaillé apparaisse, comme sorti de nulle part". Preuve que la perception de l’IA n’est pas la même que celle de l’être humain. David Gunning, qui travaille pour le programme de la DARPA, planche lui sur plusieurs pistes, comme obliger le réseau à générer un explication sémantique à chaque niveau de neurone. L’ordinateur indiquerait par exemple sur quel point de détail (couleur, moustache, yeux...) s’est appuyé le neurone pour conclure qu’il

reconnait un chat ou un renard.

L’algorithme aussi insondable que l’être humain

Pourtant, malgré ces recherches et les voeux pieux, le Deep Learning ne sera probablement jamais en mesure d’être aussi transparent et traçable que le bon vieux code informatique. Mais est-ce vraiment un problème ? "Les psychologues ont montré que quand vous demandez à un humain de vous expliquer un choix, il va construire une justification après coup qui ne sera sans doute pas la vraie raison", a défendu Peter Norvig, le directeur de recherche de Google, lors d’une conférence à Sydney en juin 2017. Selon lui, la même approche devrait être appliquée au Deep Learning : "On pourrait entraîner un algorithme à générer une explication automatique de son résultat en fonction des données qui lui ont été fournies".

Nos propres réactions sont souvent irrationnelles, fruit d’instinct, de notre subconscient ou de nos habitudes. Accepterons-vous demain de la machine ce que nous acceptons de nous-mêmes ?

Article écrit par Céline Deluzarche