près de 90 % de bonnes réponses (contre 94 % pour les humains). Cette hausse de performance est presque entièrement imputable à l'augmentation de la taille des modèles de langage en réseau neuronal et de la quantité de leurs données d'entraînement.

Ces réseaux encore plus larges ont-ils enfin atteint un niveau de compréhension similaire au nôtre ? Encore une fois, c'est peu probable. Les résultats de WinoGrande s'accompagnent d'importantes mises en garde. Par exemple, parce que les phrases dépendent des travailleurs d'Amazon Mechanical Turk, la qualité et la cohérence de l'écriture sont assez inégales. L'IA utilisée pour filtrer les phrases « non imperméables à Google » peut avoir été trop peu sophistiquée pour repérer tous les potentiels raccourcis statistiques qu'un énorme réseau neuronal pourrait emprunter, et elle ne s'appliquait qu'à des phrases individuelles, si bien que certaines des phrases qui restaient ont fini par perdre leur « jumelle ». Une étude postérieure à ces travaux a montré que les modèles de langage en réseau neuronal testés uniquement avec des phrases jumelles, et devant répondre correctement aux deux, sont bien moins précis que les humains, ce qui indique que le résultat de 90 % vu plus tôt est moins significatif qu'il ne pouvait le paraître.

Au final, que retenir de cette saga Winograd ? La leçon principale est qu'il est souvent difficile de déterminer, à partir de leur performance lors d'un défi donné, si des systèmes d'IA comprennent véritablement le langage (ou d'autres données) qu'ils traitent. Nous savons que les réseaux neuronaux utilisent souvent des raccourcis statistiques – au lieu de vraiment faire preuve d'une compréhension semblable à celle des humains – pour obtenir de bonnes performances sur les schémas Winograd et sur d'autres bancs d'essais orientés vers une « compréhension générale du langage ».

Le nœud du problème, à mon avis, est que comprendre le langage nécessite de comprendre le monde, et notamment ce que signifie que « la voiture de sport a doublé la camionnette de la poste parce qu'elle était plus lente ». Cela suppose de savoir ce que sont des voitures de sport et des camionnettes de la poste, que des voitures peuvent se « doubler », et, à un niveau encore plus fondamental, que les véhicules sont des objets qui existent et interagissent dans le monde, conduits par des humains avec leurs propres objectifs.

Ces connaissances, nous, humains, les tenons pour acquises, mais ce n'est pas le cas des machines. Et il est peu probable que ce soit écrit explicitement dans le texte d'entraînement de n'importe quel modèle de langage. Certains chercheurs en cognition estiment que les humains, pour apprendre le langage, s'appuient sur un noyau de connaissances, prélinguistiques et innées, de l'espace, du temps et de nombreuses autres propriétés essentielles du monde. Si nous voulons que les machines maîtrisent le langage comme nous, nous devrons d'abord les doter des principes primordiaux avec lesquels nous naissons. Et pour évaluer leur niveau de compréhension, nous devrions commencer par évaluer leur capacité à saisir ces principes, ce qu'on pourrait appeler une « métaphysique infantile ».

Entraîner et évaluer des machines au niveau d'intelligence d'un nourrisson peut apparaître comme un pas de géant en arrière par rapport aux prouesses de Watson et de GPT-3. Mais si l'objectif est une compréhension authentique et digne de confiance, il se peut que ce soit le seul chemin vers des machines vraiment capables de comprendre à quoi « il » ou « elle » fait référence dans une phrase.

— L'autrice —

> Melanie Mitchell
est professeuse de complexité à l'institut de Sante Fe, au Nouveau-Mexique, aux États-Unis.

— À lire —

> Y. Elazar *et al.*, Back to square one : Artifact detection, training and common sense disentanglement in the Winograd schema, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

> B. Taira *et al.*, A pragmatic assessment of Google Translate for emergency department instructions, *Journal of General Internal Medicine*, 2021.

> P. Amsili et O. Seminck, Schémas Winograd en français : une étude statistique et comportementale, *Actes de TALN 2017*, 2017.

> T. Brown *et al.*, Language models are few-shot learners, https://arxiv.org/abs/2005.14165.

La reconnaissance automatique fondée sur l'apprentissage profond a d'étonnantes faiblesses. Exploiter ces failles permet de s'amuser, mais aussi d'améliorer les procédures d'apprentissage… ou de concevoir de nouvelles attaques malveillantes.

Un seul pixel vous manque et tout est désappris

Jean-Paul Delahaye

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.