Voici comment rendre l’intelligence artificielle plus rapide et plus écologique

L'empreinte carbone des technologies derrière l'intelligence artificielle est considérable. Des nouvelles voies existent.

Temps de lecture : minute

3 mars 2021

Cet article est republié à partir de The Conversation France

Nous déverrouillons nos téléphones grâce à la reconnaissance faciale, nos réfrigérateurs "intelligents" nous aident à gérer nos stocks de nourriture et nos voitures se conduiront bientôt toutes seules. Les objets de notre quotidien "apprennent" constamment. Mais le volume de connaissances qu’ils peuvent accumuler est limité par la technologie actuelle. Ils ont besoin de nouveaux "neurones" plus performants et moins énergivores. La science est en voie de les trouver. Les applications qui permettent l’apprentissage automatique (*Machine Learning*), à la base de l’intelligence artificielle, sont soutenues par des réseaux de neurones artificiels (RNA).

Les RNA sont des ensembles organisés de neurones artificiels interconnectés. Ils sont créés dans le but de pouvoir effectuer des opérations complexes ou de résoudre des problèmes difficiles grâce à un mécanisme d’apprentissage semblable au fonctionnement du cerveau.

Un goulot d'étranglement

Ces réseaux ont contribué à l’avènement de l’Internet des objets ou

"objets connectés" et ils ont révolutionné la manière dont nous consommons de nombreux services dans le secteur financier, les transports, les télécommunications et les soins de santé. De nos jours, les réseaux de neurones artificiels sont principalement conçus grâce à des logiciels exigeant une puissance de calcul de plus en plus grande. Cette puissance est fournie par de gigantesques serveurs, qui sont très énergivores et dont l'empreinte carbone est considérable.

De plus, le nombre de composants électroniques qu'il est possible d'intégrer dans les réseaux de neurones artificiels arrivera bientôt à saturation et limitera les possibilités d'améliorer leur performance. Ce phénomène, connu sous le nom de "goulot d'étranglement", engendre des délais de transmission non négligeables, ce qui nuit aux services mobiles en temps réel.

À l'Institut national de la recherche scientifique – Énergie Matériaux Télécommunications (INRS-EMT), le groupe de recherche sur la photonique non linéaire, que je dirige, tente de mettre au point des micros dispositifs photoniques (qui utilisent des photons plutôt que des électrons) intelligents, et qui consomment peu d'énergie, afin d'augmenter la capacité des réseaux de neurones artificiels actuels.

Exploiter la lumière

De tels dispositifs, alimentés par des algorithmes d'apprentissage, sont faits de composants photoniques intégrés, qui exploitent les propriétés intrinsèques de la lumière pour atteindre des performances (surtout la vitesse) extrêmement élevées, avec une empreinte environnementale réduite.

Dans une étude récente publiée dans la revue *Nature* en collaboration avec le professeur David J. Moss, directeur du Centre des sciences optiques, à l'Université de technologie de Swinburne, nous avons testé un

réseau très puissant de neurones artificiels appelé réseau neuronal convolutif (RNC). Ce type de réseau peut effectuer 10 billions d'opérations par seconde.

Les réseaux neuronaux convolutifs fonctionnent sur le même principe que le cortex visuel des mammifères. Ils permettent de détecter la présence de motifs simples dans une image, comme la forme ou la couleur, et d'identifier progressivement le contenu de l'image en entier par association et recoupement. Ce type de réseaux est utilisé en particulier pour les applications de reconnaissance faciale et d'image par ordinateur, mais aussi pour la reconnaissance vocale et le diagnostic médical.

Les réseaux convolutifs décomposent une image en caractéristiques dominantes, telles que les bords, les couleurs, l'orientation du dégradé (appelé le "filtrage"), qui sont plus faciles à traiter. Ensuite, ces caractéristiques sont attribuées à des couches individuelles du réseau (appelées couches cachées), accélérant l'acquisition d'une image par étapes de "convolution".

Une avancée pour les véhicules autonomes

Mes collaborateurs et moi avons développé un réseau de neurones convolutif capable de traiter des images allant jusqu'à 250 000 pixels, à une vitesse suffisamment grande pour les applications de reconnaissance faciale. Des analyses comparatives ont montré une reconnaissance réussie des images numériques avec une précision de 88 %. Pour atteindre ce résultat, notre réseau a utilisé une couche de 10 neurones entièrement connectés et que nous avions déjà testée auparavant.

Le caractère évolutif de ce dispositif et sa compatibilité avec le matériel électronique standard offrent des perspectives intéressantes dans l'apprentissage de données massives pour des applications en temps réel et à très haut débit, comme les véhicules autonomes et la reconnaissance

vidéo en temps réel.

Comme le cerveau humain

Mon équipe et moi développons nos dispositifs en nous basant sur des architectures de réseaux de neurones récurrents. Ces réseaux, qui s’inspirent des circuits du cerveau (composés de neurones et de synapses), possèdent une mémoire "à court terme" essentielle pour traiter des séquences dynamiques de données (par exemple, pour améliorer la performance des canaux de télécommunications).

Un aspect fondamental de ce système est que le nombre de neurones nécessaires dans le réseau est moins élevé grâce au multiplexage temporel (distribution d’impulsions de temps multiples dans différents canaux). Cela nous permet de créer des neurones virtuels à partir d’un seul neurone physique. Le nombre de neurones virtuels créés varie selon les applications.

Cette technique a l’avantage de réduire la complexité et le nombre de composantes nécessaires par rapport à d’autres réseaux de neurones artificiels lors de leur conception. Les réseaux de neurones récurrents photoniques étudiés à l’INRS peuvent potentiellement accomplir un plus large éventail de tâches d’apprentissage complexes et qui exigent du temps comme la reconnaissance vocale, les prévisions financières et le diagnostic médical.

Mon équipe de recherche s’est engagée à faire progresser l’état des connaissances sur les réseaux neuronaux artificiels pour l’avènement de technologies émergentes, comme la 6G, qui nécessiteront la transmission et le traitement de grands débits de données, à une vitesse ultra-élevée, ce qui est inconcevable avec les réseaux actuels. Et le dernier impact de nos recherches, et non le moindre, est qu’elles permettront de diminuer considérablement l’empreinte carbone des applications d’apprentissage

profond et leur effet sur l’environnement.

Roberto Morandotti, integrated, nonlinear and quantum optics, Institut national de la recherche scientifique (INRS)

Article écrit par The Conversation France