

Les algorithmes d'intelligence artificielle ont fait des progrès considérables ces dernières années. Ils sont de plus en plus présents dans notre quotidien et sont utilisés pour analyser le contenu de textes ou d'images, pour reconnaître des instructions vocales, et souvent avec une efficacité supérieure aux humains. Ainsi, en 2017, le programme *AlphaGo* a battu le champion du monde coréen Ke Jie au jeu de go, un exploit qu'aucun expert n'attendait avant au moins dix ans.

Cependant, pour parvenir à de telles performances, ces algorithmes consomment des quantités très élevées d'électricité, beaucoup plus que notre cerveau pour des tâches équivalentes. Par exemple, les modèles d'intelligence artificielle de traitement du langage humain, comme BERT (ou sa version française FlauBERT) nécessitent une énergie de mille kilowattheures pour être entraînés. C'est autant que consomme le cerveau humain en six ans !

L'intelligence artificielle moderne, qui s'appuie sur des algorithmes réalisant des réseaux de neurones artificiels dits « profonds », s'inspire pourtant partiellement de la structure du cerveau (*voir l'encadré ci-contre*). Pourquoi donc un tel contraste dans les performances ? Les programmes d'intelligence artificielle tournent aujourd'hui sur des ordinateurs classiques, ou sur leurs cousins les cartes graphiques, dont l'architecture et le principe de calcul sont bien différents de ceux du cerveau.

Pour résoudre ce défi énergétique, de nombreux chercheurs et industriels développent de nouvelles approches visant une électronique dont le fonctionnement serait vraiment calqué sur celui du cerveau. C'est le cas de la spintronique, où l'on exploite certaines propriétés quantiques des électrons. Des dispositifs reposant sur cette technique sont capables de simuler le comportement des neurones du cerveau et des synapses qui les relient. Nous sommes encore loin du stade de l'industrialisation d'une solution complète fondée sur la spintronique, mais les premiers résultats sont déjà très prometteurs.

UNE SÉPARATION CALCUL-MÉMOIRE ÉNERGIVORE

Dans les ordinateurs actuels, l'obstacle majeur pour économiser de l'énergie et gagner en vitesse est lié au fait que les informations sont stockées dans une mémoire physiquement séparée des processeurs, où les données sont transformées en suivant les instructions du programme. Or les algorithmes d'intelligence artificielle comportent des millions, voire des milliards, de paramètres. Lorsqu'un tel programme opère, il doit sans cesse chercher des informations en mémoire puis les additionner ou les multiplier dans le processeur, avant de

les replacer dans la zone de stockage. Ce flux permanent de données par une voie de communication unique consomme une quantité d'électricité considérable.

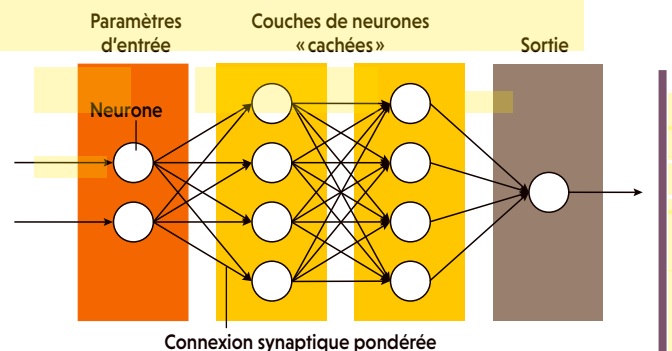
Dans le cerveau, le fonctionnement est très différent. La mémoire est partout : elle est distribuée et stockée dans les synapses, les connexions entre les neurones qui, eux, exécutent des opérations sur les données. Chaque neurone est connecté en moyenne à dix mille synapses, ce qui permet notamment de réaliser des calculs en parallèle. Cette juxtaposition entre calcul et mémoire, doublée d'un traitement parallèle, est d'une redoutable efficacité énergétique.

La nature même du calcul est une autre source de gain considérable d'énergie. Nos ordinateurs manipulent des données très précises, chacune étant codée sous la forme de bits d'information. Les opérations sur ces données

LES ALGORITHMES DE RÉSEAUX DE NEURONES

Les « réseaux de neurones » de l'intelligence artificielle, en particulier ceux dits « d'apprentissage profond », ne sont pas des structures physiques, mais des algorithmes dont le principe s'inspire du fonctionnement du cerveau. Ils représentent des nœuds, les neurones, reliés entre eux par des synapses. Dans le cas de l'apprentissage profond, les neurones sont organisés en couches. Le nombre de couches et de neurones par couche dépend de la tâche à accomplir par le réseau : analyse d'images, de textes, de sons... Le signal d'entrée est envoyé aux neurones de la première couche, qui transmettent un signal à ceux de la suivante, etc.

Au niveau de chaque neurone, les signaux issus de la couche précédente sont additionnés suivant une somme pondérée par les coefficients portés par les synapses, coefficients qui définissent la « fonction synaptique ». Cette dernière est configurée grâce à une phase d'apprentissage du réseau de neurones. On soumet des données en entrée du réseau, qui effectue un calcul et compare la solution obtenue à celle qu'on lui fournit. Prenons par exemple un réseau de neurones formé à reconnaître des images de chats. Si, pour une photo donnée, le réseau se trompe, les pondérations synaptiques sont corrigées par une méthode mathématique nommée « rétropropagation du gradient ». Celle-ci s'appuie sur des outils statistiques pour calculer la variation d'une erreur au niveau d'un neurone, de la dernière couche vers la première. Après un certain nombre d'itérations, la fonction synaptique est alors correctement configurée pour que le réseau de neurones reconnaisse les images de chats avec un taux d'erreurs très faible.



se font suivant des opérations logiques d'une rigueur implacable. Cette précision, assurément cruciale pour de nombreuses applications comme envoyer une fusée dans l'espace, ne l'est pas pour d'autres, comme reconnaître un visage familial. Or le fait d'imposer à un circuit électronique un comportement parfait où l'aléa n'est pas toléré a un coût élevé en énergie, car il faut éliminer tous les bruits parasites et corriger toutes les erreurs.

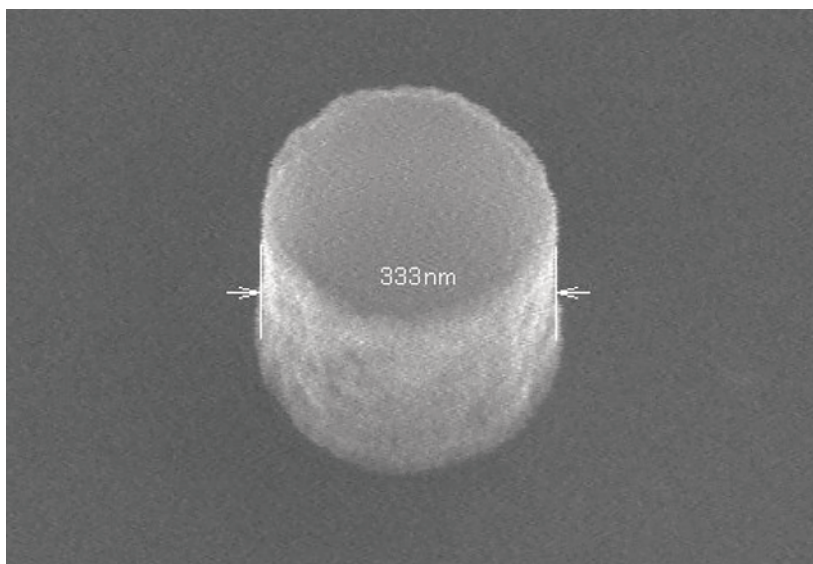
Au cours de l'évolution, le cerveau, qui ne dispose que d'une source d'énergie limitée – nos apports caloriques quotidiens – a développé une tout autre stratégie. Il fonctionne sur la base d'un compromis entre la fiabilité du traitement de l'information et la dépense énergétique. Les synapses et les neurones ont ainsi des comportements stochastiques: en intégrant une certaine marge d'aléatoire, ils ne donnent pas toujours un résultat identique même s'ils reçoivent des informations similaires.

En contrepartie, pour assurer tout de même un traitement fiable de l'information, le cerveau met en œuvre de multiples stratégies. Par exemple, comme chaque neurone, pris individuellement, n'est pas très fiable, plusieurs de ces cellules cérébrales sont parfois chargées de traiter le même signal. Un résultat cohérent et stable est ensuite déduit de l'ensemble de leurs réponses. La redondance est, en fin de compte, moins coûteuse en énergie qu'un système trop rigide qui corrige systématiquement les erreurs.

De plus, les signaux stockés dans les synapses et traités par les neurones ne sont pas des «0» ou des «1» (codage binaire) comme dans l'électronique actuelle. Ces signaux présentent une grande gamme de valeurs différentes et très proches: ils contiennent ainsi bien plus d'information, mais sont plus sensibles aux perturbations, autrement dit au «bruit». Enfin, contrairement aux processeurs où chaque élément opère au rythme d'une horloge unique, assurant un fonctionnement cohérent, mais limitant la vitesse du traitement de l'information, le cerveau travaille de façon asynchrone: chaque neurone émet des signaux électriques selon son propre rythme et s'en sert pour communiquer efficacement avec d'autres neurones, qui se trouvent parfois dans une zone différente du cerveau.

LE CERVEAU, SOURCE D'INSPIRATION

Ces spécificités contribuent à rendre le cerveau plus économe en énergie. Est-il possible de s'en inspirer pour construire des composants électroniques ou puces capables de traiter des données avec très peu d'énergie? Avec des neurones et des synapses électroniques installés au plus près, de façon à rapprocher mémoire et calcul, il serait possible de reproduire



Les jonctions tunnel magnétiques (telle celle-ci, vue au microscope électronique à balayage) sont fabriquées en plusieurs étapes. On fait croître les couches de matériaux l'une après l'autre par pulvérisation cathodique dans une enceinte sous vide. Ensuite, pour obtenir des jonctions d'un diamètre inférieur au micromètre, le matériau est recouvert d'une résine électrosensible et d'un masque sur lequel est dessinée la forme de la jonction. L'ensemble est soumis à une source d'électrons (lithographie) et la résine exposée est conservée. Les parties non protégées par la résine du matériau sont gravées et retirées pour ne garder que la forme de la jonction.

l'architecture équivalente à celle du réseau de neurones que l'on veut faire fonctionner. Cette idée n'est pas récente. Dès la fin des années 1980, Carver Mead, chercheur à Caltech (l'institut de technologie de Californie) et l'un des pères fondateurs de la microélectronique moderne, a inventé le concept de circuits «neuromorphiques», dont l'objectif est de mimer l'architecture neurobiologique du cerveau. Cette approche a longtemps été délaissée, car les obstacles technologiques étaient nombreux.

Le premier obstacle concerne la mémoire: les techniques de l'électronique actuelle ne permettent pas aujourd'hui d'intégrer d'importantes quantités de mémoire au plus près du calcul. En effet, les seuls circuits de mémoire qui peuvent être intégrés dans les processeurs (les circuits de mémoire vive statique) sont extrêmement coûteux, car ils utilisent une grande et précieuse surface dans le processeur. C'est pour cette raison que les processeurs n'intègrent que de très faibles capacités de stockage, la plus grande part de la mémoire étant physiquement distribuée en dehors des zones de calcul. Plusieurs pistes sont néanmoins à l'étude pour rapprocher mémoire et processeurs (voir l'encadré page 33). Parmi elles, l'électronique de spin, ou spintronique, a connu des progrès spectaculaires ces dernières années.

Cette technique n'exploite pas uniquement les électrons au travers de leur charge électrique, mais utilise aussi leur spin, une propriété purement quantique, assimilable à un moment cinétique (on peut représenter le spin d'un électron comme une petite flèche orientée). Comme le spin est responsable du moment magnétique intrinsèque de l'électron, quand une majorité d'électrons ont un spin pointant dans une direction privilégiée dans un matériau, ce dernier a une aimantation non nulle. La spintronique exploite les interactions ➤