

DURÉE LIBRE

FORMULE INTÉGRALE

Au lieu de
~~8,20 €~~
par mois

Partie réservée au service abonnement Ne rien inscrire

L'ESSENTIEL

> En exploitant le spin, une propriété quantique des électrons, il est possible de concevoir des composants électroniques aux caractéristiques inédites.

> Grâce à la spintronique, il est envisageable de placer des unités de mémoire au plus près des processeurs, à l'image

des synapses et des neurones du cerveau.

> Les premiers prototypes ont démontré que les dispositifs reposant sur la spintronique sont peu énergivores et ont des performances comparables aux algorithmes de réseaux de neurones usuels.

LES AUTEURS



JULIE GROLLIER
chercheuse à l'unité mixte de physique CNRS/Thales, à Palaiseau



DAMIEN QUERLIOZ
chercheur au Centre de nanosciences et de nanotechnologies (université Paris-Saclay/CNRS), à Palaiseau

Quand la spintronique s'inspire du cerveau

En termes d'énergie, le cerveau est beaucoup plus efficace que les ordinateurs actuels pour mémoriser, calculer ou apprendre. Est-il possible de s'inspirer de cet organe pour réaliser des dispositifs plus performants ? Oui, en s'appuyant sur le spin des électrons...

Grâce aux jonctions tunnel magnétiques, il est possible de concevoir des neurones et des synapses artificiels. Économes en énergie, ils pourraient avantageusement remplacer les algorithmes de réseaux de neurones.



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.

