

En matière 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 et mieux adaptés à l'intelligence artificielle ? Oui, en s'appuyant sur le spin des électrons...

Julie Grollier et Damien Querlioz

Les algorithmes d'intelligence artificielle ont fait des progrès considérables ces dernières années, et l'on peut les découvrir dans les articles de ce numéro. 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 1 000 kilowattheures pour être entraînés. C'est autant que consomme le cerveau humain en six ans !

L'intelligence artificielle moderne, qui s'appuie sur des réseaux de neurones artificiels dits « profonds », s'inspire pourtant partiellement de la structure du cerveau (voir les Repères, page 6). Pourquoi donc un tel contraste dans les performances ? Les programmes d'intelligence artificielle fonctionnent 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, chercheurs et industriels travaillent à une électronique dont le fonctionnement serait vraiment calqué sur celui du cerveau. C'est le cas de la spintronique, fondée sur 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. Le stade de l'industrialisation d'une solution complète fondée sur la spintronique est encore loin, mais les premiers résultats sont déjà très prometteurs.

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.

RAPPROCHER CALCUL ET MÉMOIRE

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 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 se



font selon 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 large gamme de valeurs différentes et très proches : ils

— En bref —

> En exploitant le spin, une propriété quantique des électrons, on conçoit 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.

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 rendent 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, on reproduirait l'architecture d'un 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 étaient nombreux.

Le premier obstacle concerne la mémoire : les techniques actuelles empêchent aujourd'hui d'en intégrer d'importantes quantités au plus près du calcul. En effet, les seuls circuits de